NCME national council on measurement in education

2022 ANNUAL CONFERENCE

APRIL 21-24, 2022

The Westin San Diego Gaslamp Quarter
San Diego, California

Welcome from the Program Chairs

Welcome to NCME's first-ever hybrid conference! We are so excited to have you join us for the 2022 NCME Conference occurring virtually as well as in San Diego, CA!

This year's conference theme is "Turn the Page: The Next Chapter for Educational Measurement". The conference serves as an opportunity to give serious thought to who we are, how we came to be, and who we want to be. It means revisiting our foundations, not running from them. It also means deciding what's next and doing so with diversity and inclusivity as our animating goal. In this light, we want to highlight a few of the many exciting sessions scheduled to take place.

Whether you are part of the in-person or virtual audience, there are several wonderful sessions that will take place in-person and be simulcast to a virtual audience, an NCME first. We invite you to look back at our history by attending the sessions, **What Does Philosophy and History Have to Contribute to Educational Measurement?** (Fri, April 22, 2:30 – 4:00 pm PT) and **Educational Measurement: How We Got to Now** (Fri, April 22, 8:00 – 9:30 am PT). We encourage you look toward our future with **Admissions Testing, Adverse Impact, and the Responsibility of the Testing Industry**(Sat, April 23, 9:45 – 11:15 AM PT) and **What does a Socially Responsible Future Look Like for Admission Testing?** (Sat, April 23, 11:30 – 1:00 pm PT). Moreover, we invite you to wrestle with diversity and inclusivity in the industry by attending **Investigating and Addressing Bias in Board Certification Exams** (Sun, April 24, 9:45 – 11:15 am PT), **Opportunity-to-learn as a Means to Enhance Equity and Increase Understanding** (Sun, April 24, 8:00 – 9:30 am PT), **De-Centering Whiteness in Assessment Practices and Products** (Fri, April 22, 4:15 – 5:45 pm PT) and **Amplifying the Voices of Women of Color in Educational Measurement** (Sat, April 23, 1:15 – 2:45 pm PT).

We also encourage you to explore **Methods to Investigate the Impact of COVID-19 on Assessment Results** (Fri, April 22, 9:45 to 11:15am PT) and **Best Practices in Evaluating Computer Scoring of Constructed Responses for Educational Measurement** (Fri, April 22, 11:30am to 1:00pm PT). For those in academia, consider attending **Recruiting and Retaining New Educational Measurement Faculty**

(Sun, April 24, 8:00 to 9:30am PT). Finally, we encourage all to stick around to the final day for fantastic sessions including **Psychometric Considerations in the Measurement of Social-emotional Learning and School Climate** (Sun, April 24, 9:45 to 11:15am) and **Advancing Contemporary Validity Theory and Practice: An Interactive Town Hall** (Sun, April 24, 2:30 to 4:00pm). These are among the many wonderful sessions that will be presented at the conference, with a significant number available to both an in-person and virtual audience.

This year, we will be hosting brand-new demonstration presentations. Demonstration session have been curated to focus on innovations in software (Virtual, April 9, 1:30 – 3:00 pm ET, in-person, Fri, April 22, 8:00 – 9:30 am PT; Sat, April 23, 9:45 – 11:15 am PT) and teaching (Fri, April 22, 2:30 – 4:00 pm PT), offered by a variety of presenters at various career levels, including graduate students. Although research blitzes have been around for two years now, this will be the first time they will be held at an in-person event. We are excited about the three in-person research blitz sessions (Fri, April 22, 1:15 – 2:15 pm, 4:15 – 5:45 pm; Sun, April 24, 9:45 – 11:15 am PT) as well as several virtual research blitz sessions during the virtual conference day on April 9. Further, we invite all attendees to check out the four in-person electronic poster board sessions (San Diego Ballroom) on Saturday, April 23, with one (Sat, April 23, 11:30 – 1:00 pm PT) devoted to the wonderful in-progress graduate student research presentations.

We are so thankful to the members and colleagues of NCME who have contributed to this engaging program, including those who have charitably volunteered their time and expertise. We are appreciative of the reviewers for providing helpful feedback as well as colleagues who volunteered as discussants and chairs. We want to thank Chun Wang, the Training and Professional Development Committee Chair, as well as Sergio Galarce and Scott Holcomb, chairs of the Graduate Students Issues Committee, for their work on the program. Finally, we are so grateful for the time, patience, and camaraderie of NCME President Derek Briggs as we brainstormed, planned, and re-planned the conference.

Finally, we owe a special debt of gratitude to the work of Matthew Gaertner who was instrumental in drafting the call for proposals, brainstorming the planning of the conference, and bringing us all together. We are hosting a special event in remembrance of Matt on Thursday, April 21, from 6:15 – 7:15 pm PT followed by a first-of-its-kind at NCME, Welcome Reception from 7:30 – 9:00 pm PT. All are welcome to attend both events.

We are excited for you to join us, whether in-person or virtually, as we reconvene the NCME community! Please enjoy the Conference.

Ben Domingue, Leslie Keng, and Brian Leventhal
2022 NCME Annual Conference Co-Chairs

NBME®
YOUR PARTNER IN
ASSESSMENT EXCELLENCE
▶ LEARN MORE AT NBME.ORG

Table Of Contents

General Information 4

Floor Plans 5

NCME Board of Directors 6

Proposal Reviewers. 7

Schedule at a Glance 9

VIRTUAL SESSIONS

Friday, April 8 16

Saturday, April 9 18

Sunday, April 10. 35

Monday, April 11. 36

Wednesday, April 13 37

Thursday, April 14 38

TRAINING SESSIONS

Thursday, April 21 39

ANNUAL CONFERENCE PROGRAM

Friday, April 22. 43

Saturday, April 23 66

Sunday, April 24 89

General Meeting Information

Welcome to the 2022 NCME Annual Conference in San Diego, CA!

NCME REGISTRATION & INFORMATION DESK

The NCME Registration & Information Desk is located in the California Foyer at The Westin San Diego Gaslamp Quarter. Stop by the registration desk to pick up your conference materials including your name badge and program. Stop by the information desk to ask questions about your membership, the program, and for any other questions! If you are participating in the NCME 5K Run/Walk, please stop by the desk to pick up your shirt.

The NCME Registration & Information Desk will be open the following hours:

Thursday, April 21	**7:30 AM – 5:00 PM**
Friday, April 22	**7:30 AM – 5:00 PM**
Saturday, April 23	**7:30 AM – 5:00 PM**
Sunday, April 24	**7:30 AM – 12:00 PM**

TWITTER

Share your experience at the NCME Annual Conference by using #NCME2022

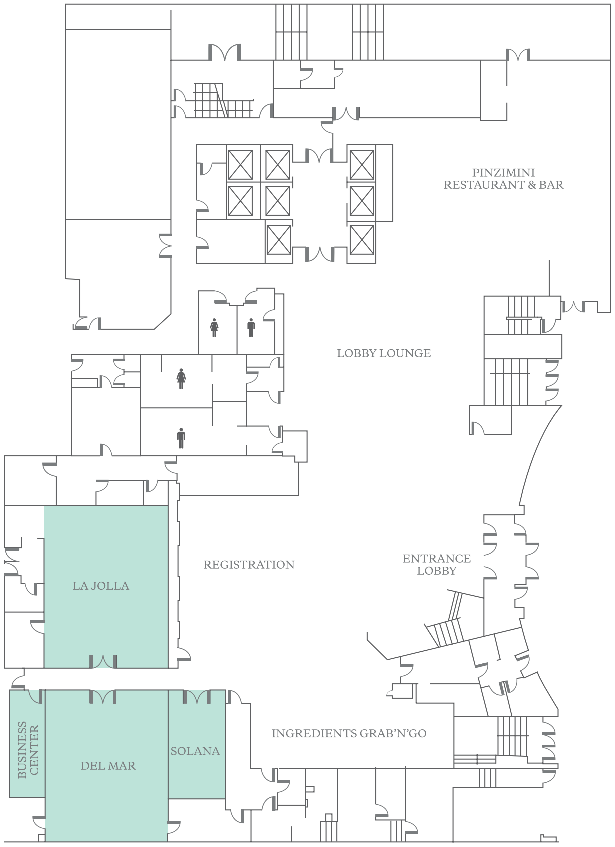
FUTURE ANNUAL CONFERENCE

2023 ANNUAL CONFERENCE

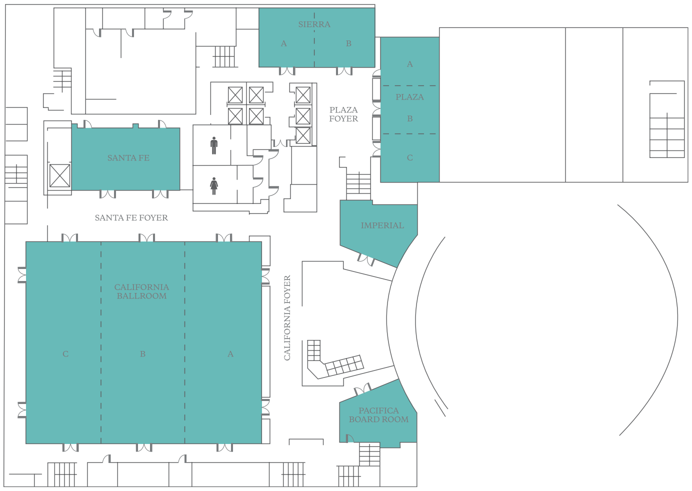
April 13-16, 2023
Chicago, Illinois

Floor Plans

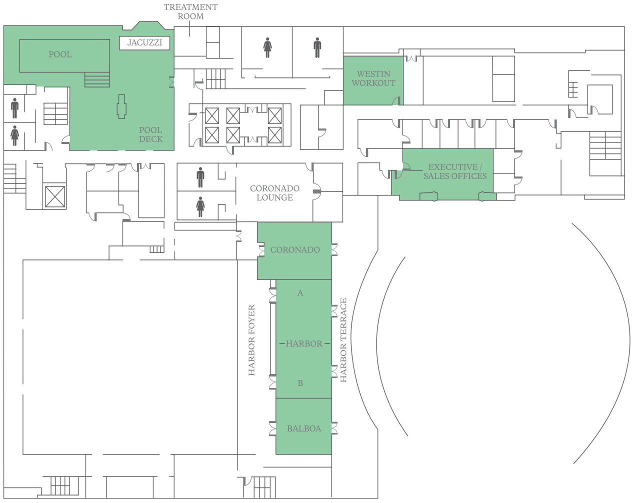
LOBBY FLOOR



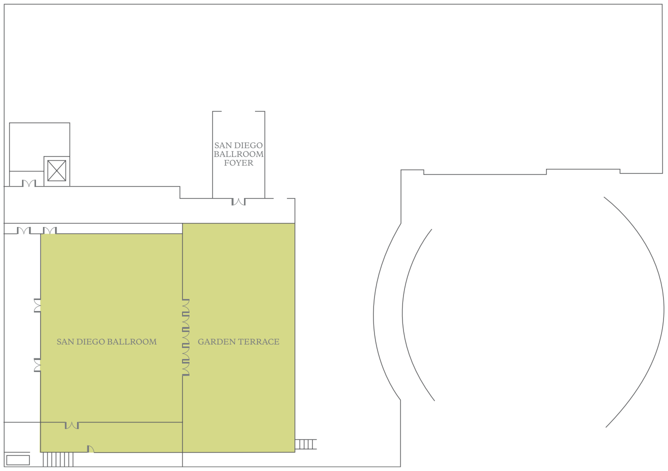
SECOND FLOOR



THIRD FLOOR



FOURTH FLOOR



NCME Board of Directors & Staff

NCME Officers

President
Derek Briggs
University of Colorado Boulder

President-Elect
Deborah Harris
University of Iowa

Past President
Ye Tong
Pearson

NCME Directors

Ellen Forte
edCount

Antionette Stroter
Chesterfield County Public Schools

Kyndra Middleton
Howard University

Sharyn Rosenberg
National Assessment Governing Board

Howard Everson
SRI International & City University of New York

Michael Walker
Educational Testing Service

Editors

Journal of Educational Measurement
Sandip Sinharay
Educational Testing Service

Holmes Finch
Ball State University

Educational Measurement: Issues and Practice
Dr. Zhongmin Cui
CFA Institute

ITEMS Editor
Brian C. Leventhal
James Madison University

NCME Book Series Editor
Kadriye Ercikan
Educational Testing Service

NCME Newsletter Editor
Art Thacker
HumRRO

NCME Website Editor
Erin Banjanovic
Pearson

2022 Annual Conference Chairs

Annual Conference Program Chairs
Ben Domingue
Stanford University

Leslie Keng
Center for Assessment

Brian Leventhal
James Madison University

Graduate Student Issues Committee Chairs
Sergio Araneda Galarce
University of Massachusetts Amherst

Scott Holcomb
University of North Carolina at Charlotte

Training and Professional Development Committee Chair
Chun Wang
University of Washington

Fitness Run/Walk Directors
Jill R. van den Heuvel, Ph.D.
Alpine Testing Solutions

Katherine Furgol Castellano, Ph.D.
Educational Testing Service

Brian F. French, Ph.D.
Washington State University

Proposal Reviewers

Anthony Albano
Benjamin Andrews
Alvaro Arce
Bozhidar Bashkov
Randy Bennett
Amy Burkhardt
Liuhan Cai
Tiago Caliço
Kevin Cappaert
Jodi Casabianca-Marshall
Hsiu-Yi Chao
Haiqin Chen
Jyun-Hong Chen
Yiling Cheng
Hye-Jeong Choi
Amy Clark
Laurie Davis
Susan Davis-Becker
Teresa Dawber
Ben Domingue
Tanzimul Ferdous
Steve Ferrara
Anthony Fina
Holmes Finch
Robert Furter
Ardeshir Geranpayeh
Raman Grover
Mihaiela Gugiu
Hongwen Guo
Qiwei He
Yong He
Ian Hembry
Tracey Hembry
Igor Himelfarb
TsungHan Ho
Thomas Hogan
Anne Corinne Huggins-Manley
Naziema Jappie
Mark Johnson
Edmund Jones
Eli Jones
Unhee Ju
Yusuf Kara
Tzur Karelitz
Leslie Keng
Justin Kern
Kyung Yong Kim
Stella Kim
Young Yee Kim
Jennifer Kobrin
Charalambos Kollias
Audra Kosh
Matthew Lavery
Brian Leventhal
Jie Li
Hwanggyu Lim
Ye Lin
Marlit Lindner
Ren Liu
Susan Lottridge
Yong Luo
Ye Ma
Jaime Malatesta
Kaiwen Man
Katie McClarty
Danette McKinley
Maria Medina-Diaz
Kristin Morrison
Katherine Nolan
Joemari Olea
Thanos Patelis
Michael Peabody
Luyao Peng
Cornelis Potgieter
Sonya Powers
Ray Reichenberg
Kelly Rewley
Michael Rodriguez
Jonathan Rollins
Jonathan Rubright
Kimberly Runyon
Kevin Santos
Edynn Sato
Christina Schneider
Robert Schwartz
Charles Secolsky
Benjamin Shear
Mark Shermis
Jessalyn Smith
Jeffrey Steedle
Dubravka Svetina Valdivia
Emily Toutkoushian
Jon Twing
Esther Ulitzsch
Aijun Wang
Jonathan Weeks
Alexander Weissman
Andrew Wiley
Phoebe Winter
Yi-Chen Wu
Adam Wyse
Hanwook Yoo
April Zenisky
Liru Zhang
Mo Zhang

Training Sessions Reviewers

Sarah Quesen

Nathan Wall

Chun Wang

Qing Yi

Graduate Student Abstract Reviewers

Ella Anghel

Sergio Araneda

Magdalen Beiting-Parrish

Masha Bertling

Maritza Casas

Roti Chakraborty

Carlos Chavez

Guanyu Chen

Onur Demirkaya

Ruiyan Gao

Guher Gorgun

Kylie Gorney

Timothy Holcomb

Jeffrey Hoover

Xuejun Ji

Radhika Kapoor

Hacer Karamese

Gamze Kartal

Russell Keglovits

Olasunkanmi Kehinde

Eunbee Kim

Nana Kim

Yun-Kyung Kim

Emma Klugman

Na Liu

Yue Mao

Kimberly McIntee

Alejandra Miranda

Aaron Myers

Monsurat Raji

Madeline Schellman

Montserrat Valdivia Medinaceli

Sarah Wellberg

Jiaying Xiao

Jiawei Xiong

Seyma Nur Yildirim-Erbasli

Yifang Zeng

Mingying Zheng

Schedule at a Glance

FRIDAY, APRIL 8

Begin Time	End Time	Location	Session Title
1:00 pm ET	5:00 pm ET	Virtual	Utility of Correspondence Analysis of Two-Way Frequency Tables for Educational Assessment*
1:00 pm ET	5:00 pm ET	Virtual	Optimal Test Design Approach to Fixed and Adaptive Test Construction using R*
1:00 pm ET	5:00 pm ET	Virtual	A Visual Introduction to Computerized Adaptive Testing*
1:00 pm ET	5:00 pm ET	Virtual	Principles and Methods in Psychometric Evaluation of Educational Assessments*
1:00 pm ET	5:00 pm ET	Virtual	Item Bank Review and Enemy Identification with Natural Language Processing*
1:00 pm ET	5:00 pm ET	Virtual	The Future of Standard Setting*

SATURDAY, APRIL 9

Begin Time	End Time	Location	Session Title
10:30 am ET	12:00 pm ET	Virtual	Addressing practical issues in assessment innovations
10:30 am ET	12:00 pm ET	Virtual	Fairness in Educational Testing: Theoretical, Research, and Practice Implications of 2014 Standards
10:30 am ET	12:00 pm ET	Virtual	Holistic Assessment and Review: Can it Increase Diversity in the Professions?
10:30 am ET	12:00 pm ET	Virtual	Handling Unusual Responses
10:30 am ET	12:00 pm ET	Virtual	Innovative Approaches in Assessment
10:30 am ET	12:00 pm ET	Virtual	GSIC Research Blitz Session
10:30 am ET	12:00 pm ET	Virtual	Challenges/Opportunities for Analytical Methods in Item Banking, Bank Evaluation, and Forecasting
12:15 pm ET	1:15 pm ET	Virtual	Advancements in Cognitive Diagnostic Modeling
12:15 pm ET	1:15 pm ET	Virtual	Applications in Adaptive Testing
12:15 pm ET	1:15 pm ET	Virtual	Fairness Topics
12:15 pm ET	1:15 pm ET	Virtual	Adaptive Testing Topics
12:15 pm ET	1:15 pm ET	Virtual	Measurement Applications in Assessment
12:15 pm ET	1:15 pm ET	Virtual	Advancements in Psychometric Techniques
12:15 pm ET	1:15 pm ET	Virtual	Measurement Methods and Applications
1:30 pm ET	3:00 pm ET	Virtual	Fairness Arguments and Implications for Born Inclusive Assessment and Validity
1:30 pm ET	3:00 pm ET	Virtual	Impact of the Pandemic from Multiple Analytic Perspectives
1:30 pm ET	3:00 pm ET	Virtual	Investigations of Bayesian Hyperprior Multigroup IRT Item Parameter Estimation Techniques
1:30 pm ET	3:00 pm ET	Virtual	Using Response Process Data to Support At-Home Administrations of High-stake Assessments

*Additional Cost

Schedule at a Glance

SATURDAY, APRIL 9

Begin Time	End Time	Location	Session Title
1:30 pm ET	3:00 pm ET	Virtual	IRT Applications
1:30 pm ET	3:00 pm ET	Virtual	Validity Topics
1:30 pm ET	3:00 pm ET	Virtual	Demonstration Session
3:15 pm ET	4:45 pm ET	Virtual	Innovation and Evolution in Standard Setting: We Can Expect Changes in Practice
3:15 pm ET	4:45 pm ET	Virtual	Let's Talk about the Solution Rather than the Problem: Recommending Changes for Educational Assessment Policy
3:15 pm ET	4:45 pm ET	Virtual	The Viability and Validity of Through-year Assessments for Instructional and Summative Uses
3:15 pm ET	4:45 pm ET	Virtual	Modeling Applications
3:15 pm ET	4:45 pm ET	Virtual	Topics in Cognitive Diagnostic Modeling
3:15 pm ET	4:45 pm ET	Virtual	Scaling, Linking, & Equating Beyond our Comfort Zones
5:00 pm ET	6:30 pm ET	Virtual	Psychometric Applications in Assessment
5:00 pm ET	6:30 pm ET	Virtual	Advancements in Measurement Approaches
5:00 pm ET	6:30 pm ET	Virtual	Psychometric and Assessment Topics
6:30 pm ET	8:00 pm ET	Virtual	Past Presidents Dinner (Invite Only)

SUNDAY, APRIL 10

Begin Time	End Time	Location	Session Title
1:00 pm ET	5:00 pm ET	Virtual	Addressing the Data Challenges from Next-generation Assessments: Data Science Upskilling for Psychometricians (Part 1) *
1:00 pm ET	5:00 pm ET	Virtual	Next-Generation Cognitive Diagnosis for Small Educational Testing Settings: Innovations and Implementation (Part 1) *
1:00 pm ET	5:00 pm ET	Virtual	Data Visualization and Analysis in the Era of COVID-19 (Part 1) *

MONDAY, APRIL 11

Begin Time	End Time	Location	Session Title
1:00 pm ET	5:00 pm ET	Virtual	Addressing the Data Challenges from Next-generation Assessments: Data Science Upskilling for Psychometricians (Part 2) *
1:00 pm ET	5:00 pm ET	Virtual	Next-Generation Cognitive Diagnosis for Small Educational Testing Settings: Innovations and Implementation (Part 2) *
1:00 pm ET	5:00 pm ET	Virtual	Data Visualization and Analysis in the Era of COVID-19 (Part 2) *

WEDNESDAY, APRIL 13

Begin Time	End Time	Location	Session Title
1:00 pm ET	5:00 pm ET	Virtual	Process Squared: Mining the processes in NAEP Process Data*

*Additional Cost

Schedule at a Glance

THURSDAY, APRIL 14

Begin Time	End Time	Location	Session Title
1:00 pm ET	5:00 pm ET	Virtual	Tools and Strategies for the Design and Evaluation of Interactive Dashboard Reports*
1:00 pm ET	5:00 pm ET	Virtual	Applications of Keystroke Logging in Educational Research*

THURSDAY, APRIL 21

Begin Time	End Time	Location	Session Title
8:00 am PT	4:30 pm PT	California Ballroom A	Building Custom Interactive Dashboards with Shiny: A Tutorial with Examples*
8:00 am PT	4:30 pm PT	Plaza A	Applications of Language models in Educational Assessment*
8:00 am PT	4:30 pm PT	California Ballroom C	Bayesian Networks in Educational Assessment (Book by Springer) *
8:00 am PT	4:30 pm PT	La Jolla	Analyzing NAEP/TIMSS Data with Direct Estimation in R, Theory and Practice*
8:00 am PT	4:30 pm PT	Plaza B	Cognitive Diagnosis Modeling: A General Framework Approach and Its Implementation in R*
8:00 am PT	4:30 pm PT	Santa Fe	Using Stan for Bayesian Psychometric Modeling
8:00 am PT	12:00 pm PT	Del Mar	Using SAS for Monte Carlo Simulation Studies in Item Response Theory*
8:00 am PT	12:00 pm PT	Sierra A	Applying Data Mining Methods to Detect Test Fraud*
8:00 am PT	12:00 pm PT	Sierra B	Computerized Multistage Testing: Theory and Applications (Book by Chapman and Hall) *
10:00 am PT	4:30 pm PT	California Ballroom B	Classroom Assessment Committee Conference Pre-session: K-12 Educator Collaboration Day
12:30 pm PT	4:30 pm PT	Del Mar	Non-commercial IRT-based Simulation Software: WinGen3, SimulCAT, MSTGen, and IRTEQ*
12:30 pm PT	4:30 pm PT	Sierra A	Sequence Mining Methods on Process Data in Large-Scale Assessments*
12:30 pm PT	4:30 pm PT	Sierra B	Embedded Alignment and Standard Setting in Practice*
6:15 pm PT	7:15 pm PT	Garden Terrace	Matthew Gaertner Remembrance
7:30 pm PT	9:00 pm PT	Garden Terrace	Welcome Reception

FRIDAY, APRIL 22

Begin Time	End Time	Location	Session Title
8:00 am PT	9:30 am PT	**Hybrid** | California Ballroom A	Educational Measurement: How We Got to Now
8:00 am PT	9:30 am PT	**Hybrid** | California Ballroom B	Various Ways of Using Context and Process Data in Measurement
8:00 am PT	9:30 am PT	California Ballroom C	Connecting COVID-19 Policy to Practice in Indiana using State Assessment Data
8:00 am PT	9:30 am PT	Del Mar	Item Innovations in CAT
8:00 am PT	9:30 am PT	La Jolla	Software Demonstrations #2

*Additional Cost

Schedule at a Glance

FRIDAY, APRIL 22

Begin Time	End Time	Location	Session Title
8:00 am PT	9:30 am PT	Santa Fe	Multidimensional Space
8:00 am PT	9:30 am PT	Plaza	Methodological Innovations in TIMSS and PIRLS: Robust Methods, Process Data, Artificial Intelligence
9:45 pm PT	11:15 am PT	**Hybrid** | California Ballroom A	Reflections on Edmund Gordon's Pedagogical Troika: Integrating Assessment, Teaching and Learning
9:45 pm PT	11:15 am PT	**Hybrid** | California Ballroom B	Methods to investigate the impact of COVID-19 on assessment results
9:45 pm PT	11:15 am PT	California Ballroom C	Considerations in the Design of Through Year Computer Adaptive Assessments
9:45 pm PT	11:15 am PT	Del Mar	Advancements in CAT
9:45 pm PT	11:15 am PT	La Jolla	Process Data Topics
9:45 pm PT	11:15 am PT	Santa Fe	Assessing Personal Skills and Qualities for High-Stakes Higher Education Admissions
9:45 pm PT	11:15 am PT	Plaza	Medical Education and Testing
11:30 am PT	1:00 pm PT	**Hybrid** | California Ballroom A	Advancing and Assessing Civic Learning to Promote Equity
11:30 am PT	1:00 pm PT	**Hybrid** | California Ballroom B	Best Practices in Evaluating Computer Scoring of Constructed Responses for Educational Measurement
11:30 am PT	1:00 pm PT	California Ballroom C	Test Optional Policy: The Current Trends and Impact on Applications and Admissions
11:30 am PT	1:00 pm PT	Del Mar	Digital-transitions, Device-effects and Disadvantage in Large-Scale Assessments
11:30 am PT	1:00 pm PT	La Jolla	Test Taking Behavior/Engagement Models
11:30 am PT	1:00 pm PT	Santa Fe	Item Detection and Design
11:30 am PT	1:00 pm PT	Plaza	Growth/Longitudinal Methods
11:30 am PT	1:00 pm PT	**Manchester Grand Hyatt** Seaport Ballroom E	Addressing Differential Educational Outcomes for Marginalized Populations in the Era of COVID-19
1:15 pm PT	2:15 pm PT	**Hybrid** | California Ballroom A	Philosophical realism in educational measurement: Is there a there there? And why does it matter?
1:15 pm PT	2:15 pm PT	**Hybrid** | California Ballroom B	Research Blitz
1:15 pm PT	2:15 pm PT	California Ballroom C	Applications in Linking & Equating
1:15 pm PT	2:15 pm PT	Del Mar	Advances in Test Design/Assessment Design
1:15 pm PT	2:15 pm PT	La Jolla	Security, Proctoring and More
1:15 pm PT	2:15 pm PT	Santa Fe	DIF Approaches in CAT
1:15 pm PT	2:15 pm PT	Plaza	Survey Research
2:30 pm PT	4:00 pm PT	**Hybrid** | California Ballroom A	What does philosophy and history have to contribute to educational measurement?
2:30 pm PT	4:00 pm PT	**Hybrid** | California Ballroom B	Methods and Results in Monitoring Hybrid Automated/Hand-scoring of Essays

Schedule at a Glance

FRIDAY, APRIL 22

Begin Time	End Time	Location	Session Title
2:30 pm PT	4:00 pm PT	California Ballroom C	Interpreting COVID-19 Test Scores: Mode Effects and Missing Data
2:30 pm PT	4:00 pm PT	Del Mar	Teaching Demonstrations
2:30 pm PT	4:00 pm PT	La Jolla	Practical Considerations in Transitioning IRT Calibration Software
2:30 pm PT	4:00 pm PT	Santa Fe	Applications of Bias, Fairness, DIF
2:30 pm PT	4:00 pm PT	Plaza	Advanced Methods Using Large Scale Assessment Data
4:15 pm PT	5:45 pm PT	**Hybrid** | California Ballroom A	De-Centering Whiteness in Assessment Practices and Products
4:15 pm PT	5:45 pm PT	**Hybrid** | California Ballroom B	At the crossroads of psychometrics and causal inference
4:15 pm PT	5:45 pm PT	California Ballroom C	Applications of Response Time in Measurement
4:15 pm PT	5:45 pm PT	Del Mar	Cheating Detection: A Collaborative Case Study using IT Certification Exams
4:15 pm PT	5:45 pm PT	La Jolla	Research Blitz
4:15 pm PT	5:45 pm PT	Santa Fe	A Content-Referenced Approach to the Interpretation of Growth
4:15 pm PT	5:45 pm PT	Plaza	Connecting Ambitious Teaching and the Formative Assessment Process

SATURDAY, APRIL 23

Begin Time	End Time	Location	Session Title
6:30 am PT	7:30 am PT	Garden Terrace	NCME Yoga
8:00 am PT	9:30 am PT	**Hybrid** | California Ballroom A	Sociocultural Context of Assessment
8:00 am PT	9:30 am PT	**Hybrid** | California Ballroom B	Components of a Well Balanced Assessment System
8:00 am PT	9:30 am PT	San Diego Ballroom	eBoard Session
8:00 am PT	9:30 am PT	Del Mar	Innovative Assessment Systems: Informing the Future of Educational Measurement
8:00 am PT	9:30 am PT	La Jolla	PISA Applications
8:00 am PT	9:30 am PT	Santa Fe	COVID-19 Impact
8:00 am PT	9:30 am PT	Plaza	Focus on Equity and Social Justice
8:00 am PT	9:30 am PT	California Ballroom C	Bifactor Models
9:45 am PT	11:15 am PT	**Hybrid** | California Ballroom A	Admissions Testing, Adverse Impact, and the Responsibility of the Testing Industry
9:45 am PT	11:15 am PT	**Hybrid** | California Ballroom B	Machine Learning Topics
9:45 am PT	11:15 am PT	San Diego Ballroom	eBoard Session
9:45 am PT	11:15 am PT	Del Mar	Broken Systems of Assessment: Addressing Fairness and Equity Challenges in Special Populations
9:45 am PT	11:15 am PT	La Jolla	Software Demonstrations #1
9:45 am PT	11:15 am PT	Santa Fe	Measuring the impact of COVID: Methodological challenges in understanding unfinished learning

Schedule at a Glance

SATURDAY, APRIL 23

Begin Time	End Time	Location	Session Title
9:45 am PT	11:15 am PT	Plaza	IRT Equating and Linking
9:45 am PT	11:15 am PT	California Ballroom C	Improving the Accuracy of Aggregate Growth Measures: Statistical Methods and Practical Challenges
11:30 am PT	1:00 pm PT	**Hybrid** | California Ballroom A	What Does A Socially Responsible Future Look Like For Admissions Testing?
11:30 am PT	1:00 pm PT	**Hybrid** | California Ballroom B	Fit Statistics
11:30 am PT	1:00 pm PT	San Diego Ballroom	GSIC eBoard Session
11:30 am PT	1:00 pm PT	Del Mar	Students with Disabilities
11:30 am PT	1:00 pm PT	La Jolla	Automatic Item Generation: Research & Applications
11:30 am PT	1:00 pm PT	Santa Fe	AERA Division D and NCME: Building on our Synergies and Singularities
11:30 am PT	1:00 pm PT	Plaza	Classroom Assessment and Data Literacy
11:30 am PT	1:00 pm PT	California Ballroom C	Handling Rapid Guessing
1:15 pm PT	2:45 pm PT	**Hybrid** | California Ballroom A	Amplifying the Voices of Women of Color in Educational Measurement
1:15 pm PT	2:45 pm PT	**Hybrid** | California Ballroom B	Using Sequence-Based Methods on Process Data in Large-Scale Assessments
1:15 pm PT	2:45 pm PT	San Diego Ballroom	eBoard Session
1:15 pm PT	2:45 pm PT	Del Mar	Bayesian Methods
1:15 pm PT	2:45 pm PT	La Jolla	Using GitHub for Open-Source Analytics, Reporting, and Dissemination of Research
1:15 pm PT	2:45 pm PT	Santa Fe	Issues in Remote Testing During the Pandemic
1:15 pm PT	2:45 pm PT	Plaza	Improving Fairness Evaluations in Automated Scoring
1:15 pm PT	2:45 pm PT	California Ballroom C	Response Time Models
3:00 pm PT	4:30 pm PT	Garden Terrace	We're Here Too: Researchers from Historically Marginalized Groups Networking Event
4:40 pm PT	6:30 pm PT	**Hybrid** | California Ballroom A, B, C	Business Meeting and Presidential Address
6:30 pm PT	9:00 pm PT	Garden Terrace	Presidential Reception (open to all)

SUNDAY, APRIL 24

Begin Time	End Time	Location	Session Title
6:00 am PT	7:00 am PT	Lobby	NCME Fitness Run/Walk
8:00 am PT	9:30 am PT	**Hybrid** | California Ballroom A	Opportunity-to-learn as a means to enhance equity and increase understanding
8:00 am PT	9:30 am PT	**Hybrid** | California Ballroom B	Recruiting and Retaining New Educational Measurement Faculty
8:00 am PT	9:30 am PT	California Ballroom C	Reimagining Assessments: The Responsibilities is Ours!

Schedule at a Glance

SUNDAY, APRIL 24

Begin Time	End Time	Location	Session Title
8:00 am PT	9:30 am PT	Del Mar	Recent Challenges to Ensuring Score Comparability
8:00 am PT	9:30 am PT	La Jolla	Validity Issues
8:00 am PT	9:30 am PT	Santa Fe	Focus on Diagnostic Classification Models
8:00 am PT	9:30 am PT	Plaza	Linking Product and Process Evidence in Student Writing
9:45 am PT	11:15 am PT	**Hybrid** | California Ballroom A	Avoiding a Train Wreck: Working with Constituents to Rethink Equity in Assessment?
9:45 am PT	11:15 am PT	**Hybrid** | California Ballroom B	Investigating and Addressing Bias in Board Certification Exams
9:45 am PT	11:15 am PT	California Ballroom C	Psychometric Considerations in the Measurement of Social-emotional Learning and School Climate
9:45 am PT	11:15 am PT	Del Mar	Research Blitz
9:45 am PT	11:15 am PT	La Jolla	Score Resolution Rules in Constructed Response Scoring
9:45 am PT	11:15 am PT	Santa Fe	English Language Learners
9:45 am PT	11:15 am PT	Plaza	Advances in Dichotomous IRT
1:15 pm PT	2:15 pm PT	**Hybrid** | California Ballroom A	Approaches to Missing Data
1:15 pm PT	2:15 pm PT	**Hybrid** | California Ballroom B	Standard Setting
1:15 pm PT	2:15 pm PT	California Ballroom C	Focus on Multistage Testing
1:15 pm PT	2:15 pm PT	Del Mar	Growth/Longitudinal Applications
1:15 pm PT	2:15 pm PT	La Jolla	Response Time Advancements
1:15 pm PT	2:15 pm PT	Santa Fe	Analysis of Multiple Choice Items
1:15 pm PT	2:15 pm PT	Plaza	Reliability Topics
2:30 pm PT	4:00 pm PT	**Hybrid** | California Ballroom A	Advancing Contemporary Validity Theory and Practice: An Interactive Town Hall
2:30 pm PT	4:00 pm PT	**Hybrid** | California Ballroom B	Diagnostic Measurement in Action
2:30 pm PT	4:00 pm PT	California Ballroom C	Innovative Methods
2:30 pm PT	4:00 pm PT	Del Mar	Assessment Issues in Competency-Based Education and Micro-Credentialing
2:30 pm PT	4:00 pm PT	La Jolla	Model-based Approach to Oral Reading Fluency Assessment
2:30 pm PT	4:00 pm PT	Santa Fe	Analysis of Polytomous Item
2:30 pm PT	4:00 pm PT	Plaza	Advances in DIF

Virtual Sessions

001. Utility of Correspondence Analysis of Two-Way Frequency Tables for Educational Assessment

Training Session
1:00 to 5:00 pm ET
Pathable: Virtual 1

This training session aims to demonstrate the utility of correspondence analysis (CA) of a two-way frequency table for educational assessment. Assuming there is a 10 x 6 frequency table where 10 rows represent school districts and 6 columns represent math and reading achievement categories (basic, proficient, and advanced), correspondence analysis can be used to examine specific associations between a certain school district and achievement categories to assess math and reading achievement in the school district. The traditional chi-squared test cannot be used because the school districts are nested within math and reading achievement categories and thus, the categories are related, violating the category independence assumption required for chi-squared testing. In addition, the chi-squared test is not designed to estimate specific associations and cannot be used for assessment. However, CA can be applied to related categories and estimate specific associations for assessment. The training will demonstrate the utility of CA for educational assessment with real achievement data from New York City School districts. This course requires basic knowledge of Statistics (e.g., knowledge about chi-squared test) and R, a laptop installed with R, and is intended for graduate students and professionals in industry and academia.

Presenter:
Se-Kang Kim, Fordham University

002. Optimal Test Design Approach to Fixed and Adaptive Test Construction using R

Training Session
1:00 to 5:00 pm ET
Pathable: Virtual 2

In recent years, fixed test forms and computerized adaptive testing (CAT) forms coexist in many testing programs and are often used interchangeably on the premise that both formats meet the same test specifications. In conventional CAT, however, items are selected through computer algorithms to meet mostly statistical criteria along with other content-related and practical requirements, whereas fixed forms are often created by test development staff using iterative review processes and more holistic criteria. The optimal test design framework can provide an integrated solution for creating test forms in various configurations and formats, conforming to the same specifications and requirements. This workshop will present some foundational principles of the optimal test design approach and their applications in fixed and adaptive test construction. Practical examples will be provided along with an R package for creating and evaluating various fixed and adaptive test formats.

Presenter:
Seung w. Choi, University of Texas at Austin

003. A Visual Introduction to Computerized Adaptive Testing

Training Session
1:00 to 5:00 pm ET
Pathable: Virtual 3

The training will provide the essential background information on operational computerized adaptive testing (CAT) with an emphasis on CAT components (including ability estimation, item exposure control and content balancing methods--weighted penalty model and shadow tests) and CAT simulation. Besides the traditional presentation through slides, this training consists of hands-on demonstrations of several key concepts, with visual and interactive tools and a CAT simulator. Practitioners, researchers, and students are invited to participate. A background in IRT is recommended. Participants should bring their own laptops and item pools in CSV file format, as they will access the tools that were designed to help the participants understand important CAT concepts and visualize the results. The tools will be available online during the workshops, and installation instructions will be provided for users wishing to install and run on their own laptops. Upon completion of the workshop, participants are expected to have 1) a broader picture about CAT; 2) a deeper understanding of the fundamental CAT techniques; 3) appreciation of the visual techniques used to analyze and present results in an intuitive and pleasing way.

Presenters:
Yuehmei Chien
David Shin, Pearson

Virtual Sessions

004. Principles and Methods in Psychometric Evaluation of Educational Assessments

Training Session
1:00 to 5:00 pm ET
Pathable: Virtual 4

Dr. Louis Roussos, with 15 years of experience in evaluating tests, will share "tricks of the trade" he has learned, including guiding principles, methods flowing from these principles, and a variety of real-life examples. Special consideration will be given to importance of communication and complex decision-making. Session will entail a mixture of lecture, dialogic learning through interactive discussion and sharing of experiences by participants, methods demonstration, and practical exercises in which participants implement the principles and methods. Exercises will result in constructive feedback to presenters and participants. Participants will: 1) understand how psychometric evaluations fit into the broader assessment cycle; 2) understand the components of comprehensive IRT-based psychometric evaluation, including a) identifying possible areas of improvement; b) making recommendations to address such improvement; c) drafting communication summary documents; and d) providing psychometric approval decisions; and 3) demonstrate their understanding by evaluating actual test-construction results and the corresponding communications. Intended audience include psychometricians and content specialists employed by testing industry vendors, clients of such vendors (e.g., department of education measurement officials), and graduate students with basic knowledge in IRT. Participants should bring a laptop with MS Word and Excel installed.

Presenters:
Louis Roussos, Cognia
Liuhan Cai, Cognia
Han Yi Kim, ACT

005. Item Bank Review and Enemy Identification with Natural Language Processing

Training Session
1:00 to 5:00 pm ET
Pathable: Virtual 5

The field of Natural Language Processing (NLP) within computer science has developed methods for indexing, categorizing, summarizing, and interpreting large numbers of text documents. The testing industry works predominantly with large collections of text (test items), and applications of NLP for developing and evaluating item banks have become more common over the past decade. This workshop will provide practical instruction on the application of NLP to item banks to identify sets of enemy items. While this is a very specific application, the methods introduced can and have been applied to a wide range of applications in the testing industry. Participants will be introduced to concepts of test similarity, topics models, and generation of semantic spaces, provided experience conducting these analyses, shown how to apply these approaches for enemy item identification, and provided ideas for additional applications in test development. Participants are encouraged to have R and Python on their computers to follow along.

Presenters:
Kirk Becker, Pearson
Fang Peng, National Council of State Boards of Nursing
J. B. Weir, National Commission on Certification of Physician Assistants

006. The Future of Standard Setting

Training Session
1:00 to 5:00 pm ET
Pathable: Virtual 6

The establishment and evaluation of cut scores on educational tests has matured as the need for accountability in high-stakes testing and public scrutiny of cut scores and impact have increased. With increasing reliance on educational tests for accountability purposes at the state, national and international level, the need for defensible standard setting methods and practices is critical. This half-day presession takes practitioners and students of standard setting from its theoretical foundations in the 1950s, through its formalization in the 1970s and proliferation of methods in the 1990s, to the theory and practice of standard setting of today. It culminates with projections of standard setting techniques and ancillary activities in the near future as well as cautions and recommendations for policy and practice. The session will include both lecture and interactive discussion and is intended for practitioners as well as researchers and graduate students. Although its primary objective is to make participants aware of factors that will influence the design and conduct of future standard settings, this session will also encourage participants to share their standard setting experiences and ideas for the future.

Presenters:
Michael Brannen Bunch, Measurement Incorporated
Daniel Bowen, Measurement Incorporated

Virtual Sessions

007. Addressing practical issues in assessment innovations

Coordinated Paper Session
10:30 to 12:00 pm ET
Pathable: Virtual 1

To support innovations in digital-based assessments, practitioners need to improve existing psychometric practices and/or develop new ways for evaluating new item types, test designs, analyzing new data sources (such as log data and process data) to understand student performance and score meaning. In the set of four papers, different approaches are used to address a few situations for innovative assessments. The first paper uses cluster analysis and regression methods to investigate test taking differences between testing remotely at home and at a test center. The second seeks to improve the current item analysis (IA) procedures to provide better feedback on innovative items (rather than simple multiple-choice items) for test development and scoring. The third paper proposes and compares different alternatives, for items taken by different ability groups under a multi-stage test design, to produce comparable item statistics, derived in either classical test theory (CTT) or item response theory (IRT). The fourth paper investigates ways to automatically flag problematic item for efficient psychometric review procedures.

Session Organizer:
Hongwen Guo, *Educational Testing Service*

Chair:
Gautam Puhan, ETS

Participants:

Assessment Engagement with Remote Testing using Log Data
Hongwen Guo, Educational Testing Service

Item Analysis of Innovative Items
Han-Hui Por, ETS; Gautam Puhan, ETS; John Bonett, ETS; Chris Kelbaugh, ETS; Lei Liu, Educational Testing Service

Item Analysis with Multistage Testing Data
Ru Lu, Educational Testing Service; Paul Adrian Jewsbury, Educational Testing Service; Hongwen Guo, Educational Testing Service

Predictive Models for Item Analysis
Zhuangzhuang Han, ETS; Chen Li, ETS; Qi Diao, ETS

Discussant:
Venessa Manna, Educational Testing Service

008. Fairness in Educational Testing: Theoretical, Research, and Practice Implications of 2014 Standards

Coordinated Paper Session
10:30 to 12:00 pm ET
Pathable: Virtual 2

This coordinated panel discussion will provide a sampling of viewpoints from individuals within and outside the NCME community regarding fairness issues that arise for different individuals in educational contexts. The panelists include several of the contributors to a forthcoming book that examines the implications of fairness in testing guidelines in the 2014 Standards for Educational and Psychological Testing on theory, research, practice, and policy. The presentations will share cross-cutting themes about fairness that arose across chapters in the book and then highlight the scholarship of psychometricians and school psychologists that thoughtfully addresses fairness issues for second language learners, students needing accommodations, and use of technology-enabled assessment, and through the lens of social justice. The goals for the session are to provide multiple perspectives on fairness that highlight complexity and variation in meanings; use scholarship, best practice methodologies, and/or examples of practice for a contextualized treatment of fairness; inform next-generation research and practice as well as future revisions to the Standards, and prompt the reevaluation of assumptions about what fairness is and methodologies for addressing it.

Session Organizer:
Jessica L. Jonson, Buros Center for Testing-UNL

Participants:

Cross-Cutting Themes about the Future of Fairness in Testing
Jessica L. Jonson, Buros Center for Testing-UNL

Theoretical, Empirical, and Practical Fairness Issues in Testing English Learners
Samuel O Ortiz, St John's University

Fairness and Assessment Using Test Adaptations
Ryan Kettler, Rutgers, The State University

Fairness Concerns Resulting from Innovations and Applications of Technology to Assessment
Wayne J. Camara, LSAC

Social Justice and Fairness
Gregory Camilli, Law School Admission Council

009. Holistic Assessment and Review: Can it Increase Diversity in the Professions?

Coordinated Paper Session
10:30 to 12:00 pm ET
Pathable: Virtual 3

Many professions (e.g., accounting, dentistry) have a documented shortage of practitioners who represent minority populations. This session examines efforts to increase diversity in professional education and practice, focusing primarily on holistic review, but also considering other ways that the assessment community contributes to efforts to increase diversity. This session brings together individuals that represent different professions and that speak to the role of assessment at different stages in the professional preparation pipeline. The professions include law, medicine, and teaching, while the stages in the pipeline include admissions testing, admissions review and student selection, certification, and licensure testing. Speakers address topics such as the role of certification standards in promoting diversity; holistic review in medicine and law; the use of situational judgment tests (SJTs) in increasing diversity; the role of traditional admissions tests in holistic review; and how principles of holistic review might contribute to licensure. The presenters have a breadth of experience in professions education and credentialing, while the discussants are well known for their contributions to validity theory and high-stakes testing.

Session Organizer:
Danette Waller McKinley, National Conference of Bar Examiners

Participants:

Ensuring Diversity in Certification Standards
Carol Ezzelle, National Board for Professional Teaching Standards

Implementing and Evaluating Holistic Review for Medical School Admissions
Fern Juster, McMaster University

Avoiding Ambushes on the Road to Diversity: Situational Judgment Test as Exemplar
Kelly Dore, McMaster University; Harold Reiter, McMaster University

The Role of Assessments in Advancing Educational Equity
Lily Knezevich, Law School Admission Council; Angela Winfield, Law School Admission Council

Increasing Diversity in Licensure through Optimal Pass-Fail Decision Models
Mark Raymond, National Conference of Bar Examiners

Discussants:
Suzanne Lane, University Of Pittsburgh
Michael Kane, Educational Testing Service

010. Handling Unusual Responses

Paper Session
10:30 to 12:00 pm ET
Pathable: Virtual 4

Participants:

A CPA procedure to detect item preknowledge
Onur Demirkaya, University Of Illinois At Urbana-Champaign; Jinming Zhang, University of Illinois at Urbana-Champaign
We proposed to use item exposure rates to reorder data along with three modified likelihood-based and residual-based statistics relying on both response and response time information to detect item preknowledge with a change point analysis. A simulation and empirical analyses were conducted to investigate the performances of the statistics.

Clustering algorithm comparisons for collusion detection using mixed-type data
Soo Ingrisone, Pearson; James Ingrisone, Pearson VUE
Cluster analysis is challenging using mixed-type data, which is common in testing data. The performance of five clustering methods under 12 different simulated conditions are examined and then applied to real data. This study provides guidance in selecting optimal clustering strategies for detecting aberrant examinees in the mixed-type data context.

Designing for Security: How Many Chameleon Clones Are Necessary to Detect Preknowledge?
Sarah Linnea Toton, Caveon Test Security; Brooke Houck, NBEO
Chameleon clones are item types designed to detect preknowledge that can be embedded in exam forms relatively quickly, making them ideal when testing organizations need to quickly bolster exam security. This research provides evidence around the necessary number of chameleon clones for the detection of preknowledge.

Iterative Detection and Treatment of Abnormal Group Level Responses
Yishan Ding, University Of Maryland; Ji Seung Yang, University of Maryland
The current study examines the performance of group fit statistics when the item parameters are biased due to the calibration sample being compromised by aberrant response behaviors. It also explores the iterative data cleansing approach in applying multilevel IRT models to improve the performance of group fit statistics.

Partitioning Variability in Test-Taking Effort and Test Emotions During a Low-Stakes Test

Beth Perkins, James Madison University; Dena Pastor, James Madison University; Sara Finney, James Madison University

We used multilevel modeling to examine whether fluctuations in effort during testing relate to fluctuations in emotions during a low-stakes institutional accountability test. The majority of variability in effort and emotions was between-examinees. Average levels of emotions were predictive of effort, whereas fluctuations in emotions during testing were not.

Discussant:

Daniel Bolt, University Of Wisconsin, Madison

011. Innovative Approaches in Assessment

Paper Session

10:30 to 12:00 pm ET

Pathable: Virtual 5

Participants:

Exploring Item-Level Revision Behaviors by Revision Log Clustering

Anqi Li; Susu Zhang, University of Illinois at Urbana-Champaign; Shiyu Wang

In this study, we propose to analyze and interpret the variable-length revision process data using recently developed process data analysis methods. Specifically, we aim at feature extraction and interpretation of the unstructured log revision data. Types of review and revision behaviors are identified and further examined across different testing conditions.

Finding students' ideas in formative science constructed response tasks

Brian Riordan; Sarah Bichler, University of California-Berkeley; Kenneth Steimel, ETS; Allison Bradford, University of California-Berkeley

As interactive online learning environments for K-12 science grow more prevalent, there is an increasing need for automated methods for detailed analysis of student written responses to provide individualized guidance. We develop and evaluate new natural language processing methods for identifying student ideas in formative constructed response science assessments.

Parent Model Calibration in the Context of Automatic Item Generation

Wei S. Schneider, The College Board; Jianshen Chen, College Board; Thomas Proctor, College Board

Obtaining accurate parent model parameters is critical as automatic item generation becomes prevalent in educational measurement. This paper proposes a two-step procedure for parent model calibration and compares it with five practical averaging methods. Results indicated that the proposed procedure is a promising and feasible approach.

Virtual Standard Setting: Which medium do judges prefer during each stage?

Charalambos Kollias, National Foundation for Educational Research

Due to limited resources, geographical distances, or a global pandemic, virtual standard setting has become more prevalent. This paper reports on the quantitative and qualitative analysis of survey items exploring whether judges preferred the audio or the video medium during a particular virtual standard setting stage.

Discussant:

Nathan Dadey, Center for Assessment

012. GSIC Research Blitz Session

Graduate Research Blitz Session

10:30 to 12:00 pm ET

Pathable: Virtual 6

Chair:

Scott Holcomb

Participants:

Analyzing Multimodal Time-Series Digital Reading-Process and Non-Time-Series Data with Recurrent Neural Networks

Matthew David Naveiras, Peabody College Of Vanderbilt; Sun-Joo Cho, Peabody College of Vanderbilt; Amanda Goodwin, Vanderbilt University; Jorge Salas, Vanderbilt University

In this study we designed and trained a recurrent neural network (RNN) to implement multimodal time-series reading-process data to demonstrate the added value of reading-process data to predicting students' posttest reading comprehension beyond non-time-series data. We present the design and implementation of an RNN for this purpose using Python.

Effect of Ability Distribution on IRT Observed Score Equating

Min Liang; Huan Liu, The University of Iowa; Won-Chan Lee, University Of Iowa

The primary goal of this study is to investigate the effect of ability distribution used when constructing a fitted number-correct score distribution on IRT observed score equating under the random groups design. A pseudo-forms analysis and a simulation study are conducted.

Evaluating the Quality of Competence-Based Assessment Instrument Using Generalized Linear Mixed Model

Xuejun Ryan Ji, The University of British Columbia; Amery Wu, University of British Columbia

This study aims to 1) explicate the necessities of multi-segments rating design for competence-based assessment with multiple raters and indicators, 2) evaluate the rating biases. The analysis is demonstrated with a Classroom Observation instrument via Generalized Linear Mixed Models. The results will be implicated on rubric revision and rater training.

New Item Selection Designs for Computerized Classification Test

Yingshi Huang; He Ren; Ping Chen, Beijing Normal University

This study proposed the novel idea of "stage adaptive" to tailor the item selection process with the decision-making requirement in each step and generated fresh insight into the existing response time selection method. Results indicate that a balanced item usage and stable test-taking times can be achieved.

Selection and Implementation of Accommodations: A Comprehensive Review

Maura O'Riordan, University of Massachusetts Amherst

Accommodations allow students to overcome barriers that prevent them from demonstrating their abilities. They are available instructionally and on assessments for students with disabilities and English learners. This paper seeks to compare the use of assessment and instructional accommodations to provide suggestions to make the processes more cohesive and effective.

Evaluating Item-Selection Rules in Fixed-Precision Between-Item Multidimensional Computerized Adaptive Testing

Yeonwho Kim; Stella Kim, University of North Carolina at Charlotte

The primary purpose of this study is to investigate item-selection methods of fixed-precision multidimensional computerized adaptive testing (MCAT) when between-item dimensionality is present. Three item-selection rules proposed by Braeken and Paap (2020) are examined under various simulation conditions to identify the most effective selection method.

013. Challenges/Opportunities for Analytical Methods in Item Banking, Bank Evaluation, and Forecasting

Organized Discussion
10:30 to 12:00 pm ET
Pathable: Virtual 7

The term item banking generally refers to the development and management of test content (the items) in some organizational structure or database (the bank). These banks carry with them the challenges that accompany managing a database whose richest data is stored as text, contains considerable amounts of metadata assigned to each element, and must support varying relationships between each element. How banks are populated, organized/codified, and queried are nontrivial concerns with significant downstream effects, effectively determining the quality of the links between how the construct is defined and operational test forms. This discussion will address some of the common challenges faced in item banking, with a focus on how traditional and novel methods and technologies can be used to improve the efficacy of item banking systems. Specifically, the challenges of defining and communicating bank health, evaluating bank health over time, and evaluation/management of SME item writers and raters, in the case of oral- and performance-based assessments, will be discussed. The discussion will focus on clearly articulating each problem, identifying complicating factors or pain points, and applicable solutions intended to better inform the practice of item/rater banking.

Session Organizer:
Robert Thomas Furter, Physician Assistant Education Association

Presenters:
Kirk Becker, Pearson
Tia Fechter, Office Of People Analytics
Joshua Goodman, NCCPA
Andrew Jones, American Board Of Surgery
Cynthia Parshall, Touchstone Consulting

014. Advancements in Cognitive Diagnostic Modeling

Paper Session
12:15 to 1:15 pm ET
Pathable: Virtual 1

Participants:

Developing a Cognitively Diagnostic Assessment on Scientific Inquiry under the G-DINA Model Framework

Ivy Mejia, University of the Philippines; Kevin Carl Santos, University of the Philippines-Diliman; Jose Pedrajita, University of the Philippines

The study develops a cognitively diagnostic assessment on scientific inquiry under the G-DINA model framework. Specifically, it identifies and validates the fine-grained skills under the specific feature of scientific inquiry. The best fitting cognitive diagnosis model and the reliability of attributes are examined as well.

Performance of Absolute Fit Indices for pG-DINA Model Under Q-matrix Misspecification

Joemari Olea, University of the Philippines; Bea Margarita Ladaga, UP Diliman; Kevin Carl Santos, University of the Philippines-Diliman

The performance of different absolute fit statistics, namely proportion correct, transformed correlation, and log-odds ratio, was investigated in detecting Q-matrix misspecifications for the pG-DINA model: under-specification, over-specification, and both under- and over-specification. The power rates and Type-I error rates were computed for cases with and without Q-matrix misspecifications, respectively.

Two New Methods of Q-matrix Validation for Cognitive Diagnosis

Li Jia; Mao Xiu Zhen, teacher

A large number of studies have shown that the accuracy of Q-matrix can affect the accuracy of items parameters and participants' diagnosis. Therefore, two new Q-matrix validation methods, namely maximum likelihood sum (MLS) and marginal maximum likelihood sum (MMLS), were proposed.

Discussant:

Youn Seon Lim, University of Cincinnati

015. Applications in Adaptive Testing

Paper Session
12:15 to 1:15 pm ET
Pathable: Virtual 2

Participants:

A Method for Measuring the Score Comparability of Computerized Adaptive Tests

Adam E Wyse, Renaissance

This study introduces a new method for measuring the score comparability of computerized adaptive tests (CATs) based on comparing conditional standard errors of measurement (CSEMs) for students that achieved the same scale scores. The effectiveness of the new method is illustrated using data from K-12 CATs.

Impact of Base Form Quality on MST Scores

Hacer Karamese; Won-Chan Lee, University of Iowa

The purpose of this simulation study is to investigate the potential impact of base form usage on MST scores. Simulations are performed to manipulate base form quality and evaluate its impact on scores.

The Impact of Missing Data on Parameter Estimation in Computerized Adaptive Testing

Xiaowen Liu; Eric Eric Loken, University of Connecticut

The current study investigates the impact of missingness on parameter recovery in computerized adaptive testing. Overall, recalibration of item parameters after an operational CAT was reasonable. Under some conditions, a subset of item discrimination parameters was estimated to be negative, resulting in a discontinuity in the ability continuum.

Discussant:

***Jordan Nelson Stoeger*,** Data Recognition Corporation

016. Fairness Topics

Paper Session
12:15 to 1:15 pm ET
Pathable: Virtual 3

Participants:

Exploring Variance with Different Levels of Data Disaggregation

Jun Li, University of Minnesota, Twin Cities; Qian Zhao, University of Minnesota, Twin Cities; Michael C. Rodriguez, University of Minnesota

With noncognitive measures and educational outcomes data from 157,757 students', we partitioned variance between and within groups based on race/ethnicity versus national origin (more specific data disaggregation). ICCs were twice or more for groups based on national origin. We discuss implications of specificity of data disaggregation for policy and practice.

Exposing Racialized Truths about the Roles of Educational Measurement & Assessment Practices

Kimberly G. L. McIntee

The purpose of this study is to illustrate how educational assessments affect marginalized communities by (mis) representing Black and Indigenous youth through testing and score reporting, and to promote new ways of test development and reporting by reformatting unjust and traditional standardized practices. Examples of more anti-racist testing practices are provided.

The Effect of Patient-Physician Gender Concordance on Item Performance

Chen Tian, University of Maryland, College Park; Kelly Rewley, American Board of Internal Medicine

We explored if physicians perform differently on medical certification exam items involving patients of the opposite gender. Examining differential item functioning (DIF) using hierarchical logistic regression, we found patient gender explains a small proportion of DIF variation. Patient-physician gender concordance is related to better item performance, especially for female patients.

Discussants:

Hongli Li, Georgia State University
Wayne J. Camara, LSAC

017. Adaptive Testing Topics

Research Blitz Session
12:15 to 1:15 pm ET
Pathable: Virtual 4

Chair:

Soo Ingrisone, Pearson

Participants:

An Evaluative Index for Examining Misrouting Impact in MST

Liuhan Cai, Cognia; Louis Roussos, Cognia

This study proposes an index to evaluate the impact of misrouting on the final measurement error on students' scores under a 1-3 MST design. This index incorporates the probability of routing from Stage 1 to Stage 2 modules and TIF for each of the stage modules.

DIF Analyses for Embedded Field Test Items in a CAT

Liu Liu, University of Washington; Aimee Boyd, Curriculum Associates

This study investigates Differential Item Functioning (DIF) for Embedded Field Test (EFT) items in the i-Ready Diagnostic Reading CAT, by varying source data by grade levels with gender and ethnicity groups. Findings revealed no difference in DIF results from using only on-grade level student responses or across grade student responses.

Effects of Consistency in Interim and Final Scoring Methods in CAT

Chunxin Wang, ACT; Yi He, ACT; Jie Li, Ascend Learning

This study investigates the effects of consistency in interim and final scoring methods for fixed-length computerized adaptive tests (CAT). Results will provide information on the impact of applying the same or different scoring methods in interim scoring and final scoring on final theta score estimation in practice.

Investigating Effect of Item Location on Item Drift in a Reading CAT

David Shin, Pearson; Yao Xiong, Pearson Assessments; Yu (Tracy) Zhao

This research is to investigate the impact of item location on item drift in reading CAT. Overall, it is observed that there are some tendencies for items to drift when the FTed position is different from the operational position. However, the amount of drift is not large.

Investigating the Effect of Item Pool Characteristics on a Fixed-Length CAT

Yi He, ACT; Chunxin Wang, ACT; Yu Su, ACT

This study investigates the effect of item pool characteristics (i.e., pool sizes, pool difficulty, and content balance) on a fixed-length computerized adaptive test (CAT). Simulation results will provide information on the adequate pool characteristics that will yield a CAT with desired measurement precision as well as item exposure.

Shortening Test Length Using an Early Termination Multistage Testing Design

Haiqin Chen, ADA; Tina Collier, American Dental Association; Matt Grady, American Dental Association

An early termination multistage testing (MST) design is proposed to shorten test lengths. This study shows that the two-stage MST allows for a reduced test length to flagged test-takers without sacrificing classification accuracy and consistency, as compared to results had these test-takers completed the full-length test.

018. Measurement Applications in Assessment

Research Blitz Session
12:15 to 1:15 pm ET
Pathable: Virtual 5

Chair:
Susan Davis-Becker, ACS Ventures, LLC

Participants:

Exploration of Selective Domain Tests in Benchmark Adaptive Assessments

Rong Jin, National Board of Osteopathic Medical Examiners, Inc.; Unhee Ju, Riverside Insights; JongPil Kim, Riverside Insights

Inclusion of domains only taught in a benchmark assessment can provide essential information for teaching and learning in the classroom with less burden for students by not taking additional domain items. This CAT simulation study explores the selective domain tests and compares their ability estimates with the original full test.

Exploring the Effect of Item Level Features in Three-Dimensional NGSS Assessments

Shuangshuang Xu; Dandan Liao, Cambium Assessment, Inc.; Frank Rijmen, Cambium Assessment, Inc

The present study focuses on key item features of items developed to assess the three-dimensional Next Generation Science Standards (NGSS) assessments. An operational item bank was analyzed to explore the relationships among response time, item difficulty, number of assertions, item meta data, and interaction subtypes.

Investigation of the COVID-19 Pandemic Impact on Achievement Skill-Level Performance

Sid Sharairi, Riverside Insights; Sharon Frey, Riverside Insights; JongPil Kim, Riverside Insights

The study compares students' performance and item/skill-domain difficulties estimated pre-pandemic vs. post-pandemic using multiple years of operational achievement data for the same grade cohort. Classical Test Theory and Rasch statistics will be compared to determine magnitude of the differences and evaluated with respect to the performance and items/skills assessed.

Item Mapping Three Ways

Sonya Powers, WestEd; Angela Bowzer, WestEd; Daniel Murphy, WestEd; Tavy Chen, Savvas Learning Company; Shirley Li, Savvas

Item mapping can be used for a variety of purposes including setting performance standards, developing performance level descriptors (PLDs), and creating diagnostic feedback. Each of these applications provides richer contextual information for score interpretation. We illustrate these three applications of item mapping using data from K-8 diagnostic assessments.

Measure Student Growth for Interim Assessments with Cumulative Blueprints

Jie Li, NCS - Pearson; Yao Xiong, Pearson Assessments; Siqi Chen, Pearson

This study presents a new design of interim assessments called cumulative blueprint and addresses the issues of measuring student growth for the new design. We show, through simulation studies and empirical data, how it distinguishes between the reasons why a score might have changed: pure learning gains vs. curriculum progression.

Using Limits of Agreement to Evaluate an Automated Scoring System

Edmund Jones, Cambridge Assessment English; Jing Xu, Cambridge Assessment English; Ardeshir Geranpayeh, University of Cambridge

Automated scoring of constructed responses is commonly evaluated by passing responses to both the automated system and human raters, and seeing how closely the two agree. We introduce a method from medical science for measuring the degree of agreement, and argue that it is more appropriate than current methods.

019. Advancements in Psychometric Techniques

Research Blitz Session
12:15 to 1:15 pm ET
Pathable: Virtual 6

Chair:
Tanesia Beverly, Google, LLC

Participants:

Confidence Intervals in the Marginalized Bayesian EM Estimation of IRT Models

Insu Paek, Florida State University; Zhongtian Lin, Cambium Assessment, Inc; R. Phillip Chalmers, York University

When item priors are used in the MML-EM estimation of IRT models, MAP and PSD for item parameters are obtained and they may be used to construct CIs for item parameters. This study examined the behaviors of the CIs constructed by MAP and PSD for a 2PL IRT model.

Evaluation of Bayesian Estimation in Small Sample Calibration

Jie Li, Ascend Learning

Maximum likelihood estimation requires large sample size to obtain estimates' desired statistical properties. This study evaluates Bayesian estimation in small sample Rasch model calibration with varying test lengths, ability distributions and choice of priors and estimators. Results from different priors and estimators are compared.

Incorporating Spatial Relationships Between Skills Into The Longitudinal DINA Model

David Arthur, Purdue University Statistics; Hua-Hua Chang, Purdue University

Accurate estimation of students' learning trajectories leads to better teaching practices and materials but requires accurate information about skill relationships. By using neural networks and the Longitudinal DINA (L-DINA) model, better information can be learned about skill relationships and learning trajectories, thus resulting in higher quality teaching.

Investigating the Effect of Differential Rapid Guessing on Population Invariance in Equating

Jiayi Deng, University of Minnesota; Joseph A. Rios, University of Minnesota

This study aimed to examine whether differential rapid guessing (RG) can lead to incorrect inferences concerning the population invariance assumption in test equating. Data were generated for two examinee subgroups administered two alternative forms of multiple-choice items. Both RG rates and rapid guesser ability characteristics were manipulated

Investigating the Impact of Threshold Reversal in Partial Credit Model

Fang Peng, National Council of State Boards of Nursing; William J Muntean, National Council of State Boards of Nursing; Shu-chuan Kao, NCSBN

This simulation study examines the impact of reversed thresholds commonly seen in polytomous items under the Partial Credit Model. In particular, it aims to explore whether threshold reversal leads to bias in the ability estimate and problematic item usage in the context of adaptive testing.

Methods for Imputing Scores on Constructed Response Items – Accuracy and Impact on Psychometric Property

Yanxuan Qu, ETS; Sandip Sinharay, Educational Testing Service

This study compares the performance of three missing-data imputation methods in terms of their accuracy and their impact on the psychometric properties (e.g., optimal weights and reliability of weighted scores) of a constructed-response test with three parts. Several recommendations for practice are provided.

020. Measurement Methods and Applications

Research Blitz Session
12:15 to 1:15 pm ET
Pathable: Virtual 7

Chair:

Karla Egan, EdMetric, LLC

Participants:

Effects of Different Content Strands on Student Achievement: Longitudinal Perspectives

Shudong Wang, NWEA; Ann Hu, NWEA; Wei He, NWEA

Time effect of instructional areas measured by content strands on student achievement are investigated based on longitudinal large scale standardized achievement tests using longitudinal and cross-section modeling methods. The results show that 1) different instructional areas had impact on student achievement across time and 2) impact diminished as time increased.

Assessing Mode Effects of At-Home Testing Without a Randomized Trial

Sooyeon Kim, ETS; Michael E. Walker, Educational Testing Service

We used real data to assess test mode effects associated with taking a test in a test center versus testing at home using remote proctoring. We used a pseudo-equivalent groups approach that uses examinee background information to construct sample weights via minimum discriminant information adjustment to reduce group inequivalence.

Methods for Differential Item Function with Automatic Item Generation

Wei S. Schneider, The College Board; Sunhee Kim, College Board

As Automatic Item Generation (AIG) becomes prevalent, it has introduced challenges to Differential Item Functioning (DIF) detection. We found that some of those challenges can be alleviated by applying empirical Bayes and extending the Mantel-Haenszel DIF method at both the item model and individual item levels.

Longitudinal Mixture Cognitive Diagnostic Models for Learning Trajectory

Qiao Lin, University of Illinois at Chicago; Yoon Soo Park, Harvard University

This study proposes a longitudinal mixture Cognitive Diagnostic Model (CDM) framework to detect differential attribute growth among different latent subgroups. Real-world data analysis identified group variations in the change of attribute mastery over time in the context of the latent subpopulation. Simulation studies provide inferences on estimation and parameter recovery.

Natural Language Processing and Adversarial Testing for Automated Scoring

Ye Lin, Ascend Learning; Chuan Sun, University of Kansas

In this study, we applied machine learning algorithms to explore the relationship between students' essay responses and scores. Multiple algorithms were applied for comparison. Adversarial examples were also considered to test the robustness of the algorithms. General recommendations for future automated scoring were provided to help inform scoring efforts.

Interpretation of Spurious Differential Item Functioning

Yongnam Kim; Seo Young Lee, Prometric LLC

Using a rigorous causal framework, this study first discusses the conceptual difference between item bias and differential item functioning (DIF). Based on it, the study presents some scenarios where a spurious DIF (i.e., DIF but not item bias) is structurally embedded and will be detected by various statistical methods.

021. Fairness Arguments and Implications for Born Inclusive Assessment and Validity

Coordinated Paper Session
1:30 to 3:00 pm ET
Pathable: Virtual 1

This coordinated session focuses on fairness arguments as vital to born inclusive assessment and thus to the validity of interpretations of assessment outcomes, decisions, and consequences for a broad range of students and student subgroups. Fairness arguments involve systematic consideration of the comparability of score-based inferences and consequences across groups within a population, as well as contexts for and assumptions related to comparability, and potential threats to fairness and validity (Davidson, 2021; Gholson, 2021; Xi, 2010). Born inclusive assessment refers to assessment that, from the beginning of its design and through its development, purposefully considers and addresses inclusion vis-a-vis a broad range of learners, their socio-cultural experiences, accessibility needs, different ways of learning, and the diverse ways they demonstrate learning (based on Diagram Center, 2019; Guzman-Orth et al., 2021). Presenters describe assessment design and development methodologies from the perspective of a fairness argument and validity. Discussion will focus on large-scale assessments for accountability.

Session Organizer:
Edynn Sato, Sato Education Consulting LLC

Participants:

Reimagining Fairness: Designing Equitable Educational Assessments
Melissa L. Gholson, Educational Testing Service (ETS)

Constructing Fairness and Validity Arguments: A Principled Approach for an Alternate ELPA
Edynn Sato, Sato Education Consulting LLC; Yun-Kyung Kim, University Of California – Los Angeles

Reimagining Differential Functioning to Better Understand Fairness and Demographic Intersectionality
Joseph A. Martineau, Educational Testing Service

Improving the Chances that Anchor Sets are Born Inclusive
Anne H. Davidson, Edmetric

Discussant:
Linda Cook, ETS

022. Impact of the Pandemic from Multiple Analytic Perspectives

Coordinated Paper Session
1:30 to 3:00 pm ET
Pathable: Virtual 2

We know that the pandemic has impacted the educational system in ways that we are just beginning to disentangle. This coordinated paper session presents a series of analyses regarding the impact of the pandemic on key features for a series of assessments. The first paper looks to evaluate the impact of the pandemic on the measurement invariance of an assessment. The second paper concentrates on the degree to which the impact of the pandemic translates to model/data misfit when IRT models are used. Using propensity score matching methods, the third study addresses how does student performance in 2021 compare to 2019 and whether specific domains of knowledge were more impacted than other by the disruption in student learning. While the third paper is focused on evaluating the impact of the pandemic over two years on student performance, the fourth paper looks specifically at the impact of the pandemic at the school level using hierarchical linear models. This coordinated paper session will look at the impact of the pandemic from different perspectives to provide context for the upcoming implementation of large-scale assessments in 2022 and beyond.

Session Organizer:
Marc W Julian, DRC

Participants:

Examining Measurement Invariance Before and After the 2020 Pandemic
Huan Wang, Data Recognition Corporation; Dong-In Kim, Data Recognition Corporation

The Impact of the Pandemic on IRT Model/Data Fit
Christie Plackner, data recognition corporation; Vince Struthers, Data Recognition Corporation

Evaluating Student Performance Amidst the Pandemic using Propensity Score Matching
Kim Hudson, Data Recognition Corporation; Joanna Tomkowicz, data recognition corporation; Wen-Ching Li, Data Recognition Corporation

The Pandemic Impact on School Performance
Dong-In Kim, Data Recognition Corporation; Aurore Phenow, Data Recognition Corporation

Discussant:
Karla Egan, EdMetric, LLC

023. Investigations of Bayesian Hyperprior Multigroup IRT Item Parameter Estimation Techniques

Coordinated Paper Session
1:30 to 3:00 pm ET
Pathable: Virtual 3

When prior distributions provide little information above that provided by the likelihood, then they are considered to be weakly informative. Therefore, the prior contains enough information to regularize the posterior distribution and keep it roughly within bounds, but without swamping the influence of the data as expressed by the likelihood function. Common Bayesian IRT item parameter estimation packages utilize this approach. However, if the distribution of item parameters is itself estimated (rather than fixed at some diffuse distribution), then there is the potential for more precise estimation of the parameters. This session will explore a Bayesian Hyperprior Multigroup IRT Item Parameter Estimation approach. Specific topics will include: (a) Practical shortcomings of existing IRT item parameter approaches when applied to small-sample language proficiency data and sparse CAT data, (b) Project research objectives, (c) Introduction to the Bayesian multigroup model and its implementation in STAN, along with some sample results, (d) Simulation results evaluating how STAN handles sparse data (e.g., the sparsity observed due to item field trial designs), and (e) Ways to speed up the STAN computation (e.g., within-chain parallelization and use of distributed computing to run the computation). This method excels when applied to sparse data matrices and small samples.

Session Organizer:
Tia Fechter, Office of People Analytics

Chair:
Tia Fechter, Office of People Analytics

Participants:

Practical Shortcomings of IRT for Small & Sparse Samples
Tia Fechter, Office of People Analytics

Project Research Objectives
Matt Trippe, Human Resources Research Organization

Bayesian Multigroup Model Daniel
Segall, DMDC

STAN Simulation Study Results
Glen Wallace, Human Resources Research Organization

Speeding up STAN computation
Ted Diaz, Human Resources Research Organization

Discussant:
Mark Reckase, Psychometric Solutions

024. Using Response Process Data to Support At-Home Administrations of High-stake Assessments

Coordinated Paper Session
1:30 to 3:00 pm ET
Pathable: Virtual 4

The COVID19 pandemic accelerated the implementation of remotely proctored high-stake assessments. Complementary to traditional psychometric analyses, clickstream process data that capture the fine-grained interaction information between test takers and items can provide rich diagnostic information for test interruption, unintended test-taking behaviors, and comparability & fairness under different test-taking conditions (modes). In this coordinated session, we gathered four presentations to show how data analytics and AI techniques joined forces with traditional psychometric analyses based on the response process data to help us gain deeper insights into the remote admins. We will share some interesting findings through our interactive dashboards based on the state-of-the-art dashboard technology for data science and discuss how this line of work could impact the future practice of remote proctoring of high-stake tests.

Session Organizer:
Jiangang Hao, Educational Testing Service

Participants:

Using Data Analytics and AI to Detect Unintended Test Taking Behaviors
Jiangang Hao, Educational Testing Service; Chen Li, ETS

Cloud-based Response Process Data Reduction and Mining
Chen Li, ETS

The Impact of Interruptions that Occurred During At-home Tests on Test Performance
Katherine Furgol Castellano, Educational Testing Service; Sandip Sinharay, Educational Testing Service

Do People Write Differently at Home vs. in Testing Centers?
Mo Zhang, Educational Testing Service; Hongwen Guo, Educational Testing Service; Chen Li, ETS

Discussant:
Bryan R. Drost, Rocky River City Schools

025. IRT Applications

Paper Session
1:30 to 3:00 pm ET
Pathable: Virtual 5

Participants:

A General Item Response Models for Complex Multiple-choice Questions
Kuan-Yu Jin, Hong Kong Examinations and Assessment Authority; Guo-Wei Sun, National Sun Yat-sen University
A complex multiple-choice question (CMCQ) is composed of multiple judgments, where the designed options indicate the only correct combination of the alternatives and some incorrect combinations. Beyond the analysis of designed options, this study aims to propose a new model to accommodate the alternatives in CMCQs.

A Multilevel Conway–Maxwell–Poisson IRT Model for Count Item Responses
Marian Strazzeri, US Food and Drug Administration; Ji Seung Yang, University of Maryland
Multilevel count data are often overdispersed across clusters but underdispersed within clusters. We propose and evaluate a mean-parameterized Conway–Maxwell–Poisson model that accommodates underdispersed, equidispersed, or overdispersed multilevel multivariate count responses to yield correct inferences about the data generation process, measurement properties, and individual levels of the target construct(s).

A Nonparametric Unfolding IRT Approach for Identifying Targeted Feedback in Writing
Ye Yuan, University of Georgia; George Engelhard, UGA
The study uses a nonparametric unfolding IRT model to explore and identify targeted feedback strategies for student writing. The targeted feedback strategies are the recommended by expert teachers as optimal for improving student achievement. Data from a state-wide writing assessment program is used to illustrate the model.

Calibration Timing and Seasonality Effect in Interim Assessments
Siqi Chen, Pearson; Yao Xiong, Pearson Assessments; Jie Lin, Pearson; David Shin, Pearson
Interim assessments with the same blueprint often experience seasonality where student ability change over time. Such seasonality affects the item calibration quality, and the lead time required for producing new test forms further complicates issues. This study investigates the calibration timing and seasonality within the interim assessment framework.

Investigating the Impact of Response Noise on Forced-Choice Item Response Theory Models
Bea Margarita Ladaga, UP Diliman; Joemari Olea, University of the Philippines; Kevin Carl Santos, University of the Philippines-Diliman
The study compares the latent parameter recovery of MOLE (respondent picks most and least endorsed options) and RANK (respondent ranks all possible options per block) questions by introducing noise representing random ranking of middle statements. Results showed that increasing noise levels generally diminished the performance of RANK under several factors.

Discussant:
Derek Briggs, University of Colorado Boulder

026. Validity Topics

Paper Session
1:30 to 3:00 pm ET
Pathable: Virtual 6

Participants:

Accuracy of Absolute Error Estimates within a G-theory SEM Framework

Walter Vispoel, University Of Iowa; Hyeri Hong; Hyeryung in Lee, University of Ioa; Guanlan Xu

We evaluated the accuracy of Jorgensen (2021)'s method for estimating variance components for absolute error using structural equation models. Results from numerous self-concept and personality measures revealed that Jorgensen's procedure yielded variance components, generalizability coefficients, and dependability coefficients equivalent to those produced by variance-component programs within R, SAS, and SPSS.

Technical Adequacy of a Computer-Delivered and Scored Literacy Universal Screener

Deni Basaraba, AMPLIFY Education

Universal screening will be critical as students return to school and will look different than it has in the past (NASP, 2020) and may include assessments that are computer-delivered and scored. Technical adequacy data indicate acceptable evidence of mCLASS Express for identifying students who may need additional instructional support.

Using extended Tukey's test for detecting nonadditivity in two-facet G-theory designs

Chih-Kai (Cary) Lin, National Council of State Boards of Nursing (NCSBN); Jinming Zhang, University of Illinois at Urbana-Champaign; Ye Ma, Amazon Web Services (AWS)

Detecting nonadditivity in data is important in G-theory applications because nonadditivity violates a fundamental G-theory model assumption, thereby posing threats to valid interpretation of analysis results. The current study builds on previous research and investigates the usefulness of the extended Tukey's test in detecting nonadditivity in two-facet designs.

Videos and Audios in Computer-Based Listening Tests: Item Characteristic and Response Time

Takahiro Terao, National Center for University

This study aims to compare regular and frame-by-frame videos and audio-only conditions in online listening tests. Results showed that item difficulties and discrimination powers were not different between conditions in most items. Response time did not differ in all three modes. Suggestions on item development in computer-based testing were obtained.

Discussant:

Jessica L. Jonson, Buros Center for Testing-UNL

027. Demonstration Session

Demonstration Session
1:30 to 3:00 pm ET
Pathable: Virtual 7

Participants:

A Shiny Application for Automatic Item Generation

Andrew Dallas, National Commission on Certification of Physician Assistants; Marcus Walker, National Commission on Certification of Physician Assistants; Joshua Goodman, NCCPA

Automatic item generation (AIG) has been theorized and operationalized as a method to produce items rapidly over the last two decades. The session will provide information to attendees how model creation and review can occur all nested within the Shiny application itself.

Re-envisioning Quantitative Information: Communicating NAEP Data to Parents and Policymakers

Elsie Lee-Robbins, University of Michigan; Tiago A. Caliço, American Institutes for Research; Juanita Hicks, AIR; Cadelle Hemphill, AIR

Novel NAEP visualizations are presented to demonstrate a framework for designing visualizations for distinct audiences with varying degrees of familiarity with data reporting. Reporting effectiveness of data visualizations is evaluated by leveraging learning objectives and assessments, providing empirical feedback on the effectiveness of designs which informs further development.

028. Innovation and Evolution in Standard Setting: We Can Expect Changes in Practice

Coordinated Paper Session
3:15 to 4:45 pm ET
Pathable: Virtual 1

In this session, we examine recent innovations in standard setting and propose that standard setting concepts, principles, and practices are merging with (a) testing policy and practice, and (b) principled assessment development approaches. Paper 1 introduces recent innovations and argues that they represent this merger. These innovations include vertical articulation, standards validation, benchmarked standard setting, embedded standard setting, virtual workshops, and psychometric and cognitive-social psychological theories for standard setting. Paper 2 examines how to embed standards throughout the principled assessment process. It examines the logic of different methods for setting cut scores – traditional, data-driven, and hybrid approaches – and how the logic of each supports intended score interpretations or does not. Paper 3 reports on two standard setting workshops conducted in two different environments using the Item Descriptor (ID) Matching Method: a face-to-face workshop in 2016, a synchronous virtual workshop in 2018. Paper 4 examines the challenges that changing conceptions and practice pose. Programs are reticent to consider innovations despite opportunities to increase operational efficiency, stakeholder support, and measurement integrity. This reticence is observed in educational and credentialing programs. A standard-setting expert will discuss the papers, focusing on the selected innovations and examining the claim about merging of standard setting practices.

Session Organizer:
***Steve Ferrara*, **Cognia

Participants:

Recent Innovations in Standard Setting
Steve Ferrara, Cognia

Embedding Standards to Preserve Intended Score Interpretation and Use
Robert Cook, Cognia; Daniel Lewis, Creative Measurement Solutions LLC

Turning the Page on Benchmarking and Standard Setting Studies
Voula Kanistra, Trinity College London

How Can We Bring these Innovations into Educational and Credentialing Standard-Setting Practice?
Susan Davis-Becker, ACS Ventures, LLC

Discussant:
***Wanda Swiggett*,** Educational Testing Service

029. Let’s Talk about the Solution Rather than the Problem: Recommending Changes for Educational Assessment Policy

Organized Discussion
3:15 to 4:45 pm ET
Pathable: Virtual 2

As an industry, we commonly bemoan the requirements of educational assessment and accountability policies as hindering the way in which we practice. Let’s assume that these policies will evolve over time. What changes could be made to such policies that would yield improved outcomes for students? How would assessments be different and in what ways would we, as a field, advocate for changes to the way in which they are being used? This will be the focus of discussion across this 3-part series at the 2022 meetings for the Association of Test Publishers (ATP), NCME, and the National Conference on Student Assessment (NCSA), with each session highlighting the membership perspectives of each sponsoring, professional organization. In this second session in the series, the panel includes speakers who serve as Technical Advisory Committee members, advising both state departments of education as well as their vendors on the implementation of assessment systems across the country. The panel will speak from their experiences serving as independent advisors, specifically with regards to the technical quality of statewide assessment systems, and provide their input on incorporating assessment best practices into educational policies that could contribute to improved outcomes for students and teachers.

Session Organizer:
Tracey Hembry, Alpine Testing Solutions, Inc.

Presenters:
Marianne Perie, Measurement in Practice, LLC
***Suzanne Lane*,** University Of Pittsburgh
Lillian Pace, KnowledgeWorks

030. The Viability and Validity of Through-year Assessments for Instructional and Summative Uses

Coordinated Paper Session
3:15 to 4:45 pm ET
Pathable: Virtual 4

This coordinated session explores the viability of through-year assessment systems for various uses including (1) informing instruction and (2) supporting summative claims about what a student knows and can do at the end of an academic school year. The varied uses of through-year assessment systems have raised questions about the criteria used to evaluate the technical quality of these assessments. Criteria related to the definition of assessment targets for status versus growth, content domain sampling, the relationship among the multiple forms or tests, and the aggregation and weighting of results, for example, have implications for an assessment’s design and interpretation of results. Presenters will describe their assessment’s design, their validity framework, and evidence supporting the technical adequacy of their through-year assessment vis-a-vis the assessment's claims. As relevant, student progress or growth will be addressed. Critical content and psychometric considerations and approaches will be highlighted. This session is intended to be useful to assessment designers and developers, researchers, psychometricians, and policymakers.

Session Organizer:
Edynn Sato, Sato Education Consulting LLC

Participants:

Design Ideas and Plans for Through-Year Assessments: Cognia
Steve Ferrara, Cognia

The Logical and Theoretical Basis for an Integrated Through Year Assessment
Garron Gianopulos, NWEA

Validation of Dynamic Learning Maps Instructionally Embedded Assessments
Amy Clark, ATLAS: University of Kansas

Toward a Through-year Alternate ELP Assessment for Instructional and Summative Purposes
Edynn Sato, Sato Education Consulting LLC; Yun-Kyung Kim, University Of California – Los Angeles; Li Cai, UCLA

Discussant:
Brian Gong, Center for Assessment

031. Modeling Applications

Paper Session
3:15 to 4:45 pm ET
Pathable: Virtual 5

Participants:

Evaluating Effects of Text-to-Speech on Mathematics Performance using Propensity Score Models
Haobai Zhang, University of Delaware; Xiaying Zheng, Young Yee Kim, Xiaoying Feng, American Institutes for Research
Text-to-Speech (TTS) is a universal design feature in digitally based assessment. This study evaluated effects of different levels of TTS use on mathematics performance among EL/SD and non-EL/SD students using NAEP mathematics assessment data for grade 4 and 8. Quantile treatment effects of TTS were also estimated across achievement distribution.

Examining Not-reached and Omitted Responses Using an Explanatory Item Response Tree Model
Minjeong Park; Amery Wu, University of British Columbia
In large-scale educational tests, it is common that students do not complete the test or omit the items. This study introduces an explanatory item response tree model to examine how these item nonresponse behaviors are related to person and item characteristics. Findings based on the PIRLS provided practical implications.

Hierarchical Linear Models for Within-Subjects Design: Power and Effect Size
Yue Liu, Sichuan Normal University; Hongyun Liu, Beijing Normal University
The current study extends linear mixed-effects models applied in educational and psychological experimental studies to the framework of multilevel models. Power, type I error and accuracy of effect size of the hierarchical linear models are investigated to provide recommendations about data analysis and experimental design for practitioners.

Improving Score Interpretation Through Timed Testing – A GLMM Analysis of Reading Components
Frank Goldhammer, DIPF | Leibniz Institute for Research and Information in Education, ZIB; Ulf Kroehne; Carolin Hahnel, DIPF | Leibniz Institute for Research and Information in Education, Centre for International Student; Johannes Naumann, University of Wuppertal; Paul De Boeck, OSU
The interpretation of test scores can be established by theory-based item properties (validation of construct interpretation). Two reading component skills tasks were administered traditionally and with item-level time limits to handle the speed-accuracy tradeoff. Item properties (e.g., word frequency) showed stronger effects on item difficulty in timed than untimed testing.

Statistical Inference for Regularized Differential Item Functioning
Weimeng Wang, University of Maryland, College Park; Yang Liu, University of Maryland, College Park; Jeffrey Harring, University of Maryland
L1 regularization has been successfully applied to detect differential item functioning (DIF). However, it is often difficult to draw statistical inference based on the regularized estimates of DIF effect sizes. We propose a decorrelated score test to detect DIF that is asymptotically valid under mild assumptions.

Discussant:
Adam E Wyse, Renaissance

032. Topics in Cognitive Diagnostic Modeling

Paper Session
3:15 to 4:45 pm ET
Pathable: Virtual 6

Participants:

A New Test Statistic for Absolute Fit Evaluation in Cognitive Diagnosis Models

Kevin Carl Santos, University of the Philippines-Diliman; Sean Oliver Escalante, University of the Philippines-Diliman; Iris Ivy Gauran, King Abdullah University of Science and Technology

We propose a new test statistic S for assessing model-data fit in cognitive diagnosis models. Simulation results revealed that it can detect model and Q-matrix misspecifications. It has controlled type I error and superior power compared to the existing absolute fit statistics even when the items are of medium quality.

An Application of Mixture Modelling for Cognitive Diagnosis

Joemari Olea, University of the Philippines; Kevin Carl Santos, University of the Philippines-Diliman

This study proposed the Mixture G-DINA model which incorporates G-DINA and mixture modeling to account for the inherent heterogeneity in the population. Extensive simulation studies were conducted to examine the performance of the model across different scenarios. Language assessment responses were analyzed to demonstrate the capability of the proposed model.

More Accurate Estimates of CDM Classification Accuracy

Rodrigo Schames Kreitchmann, Universidad Autonoma De Madrid; Jimmy de la Torre, University of Hong Kong; Miguel A. Sorrel, Universidad Autónoma de Madrid; Pablo Nájera, Autonomous University of Madrid; Francisco J. Abad, Universidad Autonoma De Madrid

Traditional reliability estimators for diagnostic models are inflated when item parameter estimates are biased. A multiple imputation procedure is proposed to obtain better empirical accuracy estimators. Simulation results show that the new procedure better estimates the true accuracy compared to traditional indices, particularly with small samples and low-quality items.

Sensitivity of Cognitive Diagnosis Models to Local Item Dependence

Youn Seon Lim, Univeristy of Cincinnati; Fangxing Bai, University of Cincinnati; Benjamin Kelcey, University of Cincinnati

With newly developed cognitive diagnosis models which accounted for local item dependencies, this study examines the consequences of local item dependencies by simulation studies under various conditions. These results are important because ignoring such dependencies may lead to inaccurate estimates of model parameters and misclassifications of examinees.

To Use or Not to Use: Necessity of Model Selections in CD

Chia-Yi Chiu, Rutgers, The State University; Jiaxi Wang, Rutgers, The State University of New Jersey

The study aims to assess the necessity of using the model fit indices in CD under some specific conditions by comparing their performance with those of the general CD models or methods. The results will provide valuable insights into how data analysis should be done in practice.

Discussant:

Ivy Mejia, University of the Philippines

033. Scaling, Linking, & Equating Beyond our Comfort Zones

Organized Discussion
3:15 to 4:45 pm ET
Pathable: Virtual 7

This session brings professionals together from various industries such as medicine, psychology, and business to discuss their unique challenges and approaches to scaling, linking, or equating (SLE). An educational measurement scholar with an expertise in SLE methods will discuss each panelist's unique approaches to solving SLE-related challenges in their respective settings. This session will also connect approaches used in these unique settings to those used in traditional educational settings and provide a framework within which scholars can expand their SLE research lens.

Session Organizer:

Jaime Malatesta, Graduate Management Admission Council

Moderators:

Stella Kim, University of North Carolina at Charlotte
Mengyao Zhang, National Conference Of Bar Examiners

Presenters:

James Ingrisone, Pearson VUE
Huijuan Meng, Amazon Web Services (AWS)

Discussants:

Neil Dorans
Tim Moses, College Board

034. Psychometric Applications in Assessment

Research Blitz Session
5:00 to 6:30 pm ET
Pathable: Virtual 1

Participants:

A Comparison of Three Methods for Dealing with Low-Quality Response in Surveys
Nivedita Bhaktha
In this study we examined approaches that can potentially help mitigate bias in analyses due to low quality responses (LQR). We present the results of a simulation study that considered three approaches of handling LQR in regression analysis – exclude LQR, weight LQR, and use LQR indicators as covariates.

Evaluation of Cognitively Diagnostic Assessment on Fraction Subtraction in the Philippine Context
Alvin Daiz Tenorio, Capiz State University; Emir Lenard Suerte Felipe Sicangco, Tarlac State University
This study evaluated Tatsuoka's cognitively diagnostic assessment on fraction subtraction in the Philippine context by identifying the attributes used by Filipino pupils in answering the items and validating those attributes by fitting reduced and saturated CDMs at test level and a combination of different reduced CDMs at the item level.

On-the-Fly Assembled Variable-Module-Length Multistage Testing
Jyun-Hong Chen, National Cheng Kung University; Hsiu-Yi Chao, National Taiwan Ocean University
The study proposed a newly innovative design, on-the-Fly Assembled Variable-Module-Length Multistage Testing (OVMST), to enhance adaptation degree by online assembling fixed-precision modules for each examinee. The results indicated that OVMST can efficiently improve trait estimation while maintaining the appealing features of multistage testing (e.g., item review for examinees).

Utilizing Real-time Test Data to Enhance Optimal Design in Computerized Adaptive Testing
Hsiu-Yi Chao, National Taiwan Ocean University; Jyun-Hong Chen, National Cheng Kung University
This study proposed an item selection method based on integer linear programing to optimize CAT's performance, which turns information from real-time test data into feasible test constraints to determine item administration. Results showed that the proposed method can efficiently improve both trait estimation precision and item pool usage in CAT.

Working Memory and Approaches to Learning in Kindergarten Predict Math Skill Development
Ruiyan Gao, University of Hong Kong
To understand the combined effects of working memory and approaches to learning (AtL) on math skill development, this study adopted data from from the ECLS-K:2011 cohort. Latent growth curve models show that working memory and AtL have small to medium positive interaction effects on math skills initial level and growth.

035. Advancements in Measurement Approaches

Research Blitz Session
5:00 to 6:30 pm ET
Pathable: Virtual 2

Chair:
Leslie Keng, Center for Assessment

Participants:

Methods Effect in a Game-Based Assessment
Ella Anghel; William A. Lorie, Center for Assessment
The current study explores methods effects within a game-based assessment (GBA) containing several mini-games measuring math and ELA skills. Using correlational analyses and CFA, we found that games ostensibly measuring the same skill but employing different game mechanics do not load on the same trait factor.

Advances in Test Adaptivity to Improve Measurement in Testlet-based CAT
Ozge Ersan, University of Minnesota-Twin Cities; Michael C. Rodriguez, University of Minnesota
A testlet is a group of items associated with a single stimulus. In testlet-based CAT, the common practice is selecting a testlet adaptively and administering all items bundled in the testlet. In this study, adaptively selecting both testlets and items within testlets is evaluated under various conditions.

Validating a Novel Data Literacy Assessment: Psychometric and Eye-Tracking Analyses
Fu Chen; Ying Cui, University of Alberta; Alina Lutsyk
In this study, we introduced a novel digital performance-based assessment for measuring data literacy skills. To validate the assessment, we analyzed students' assessment outcomes, response time, and eye-tracking data using both psychometric and eye-tracking analyses. Results suggest that the assessment is a reliable and valid tool for data literacy evaluation.

Item Pool Evaluation and Development for Longitudinal Computer Adaptive Testing
Ann Hu, NWEA; Xueming Li, NWEA
This study investigated a modification to the bin-and-union method for generating an optimal item pool with minimum size that guides item development for a large-scale longitudinal interim assessment. Simulation results showed that new items added to the empirical pool for the operational test improved the CAT performance significantly.

Unfolding Diagnostic Classification Models

Ren Liu, University of California, Merced; Haiyan Liu, University of California, Merced

The purpose of this paper is to apply the "ideal point process" tenet into the world of diagnostic classification models and propose a framework of unfolding diagnostic classification models to score item responses from surveys, questionnaires, and rating scales.

036. Psychometric and Assessment Topics

Research Blitz Session
5:00 to 6:30 pm ET
Pathable: Virtual 3

Chair:
Wei S. Schneider, The College Board

Participants:

Changing Gear During the Pandemic – National Benchmark Test (NBT) Moves Online

Naziema Jappie, University of Cape Town; Ashley Kevin Hatting-Niekerk, University of Cape Town

The sudden shift from contact to remote digital learning platforms brought about unique challenges to the South African education landscape forcing learners into an unfamiliar path of assessment. This paper will provide insights to the requirements for a well-designed digital assessment system which needs to be relevant, adaptable and trustworthy.

Rapid Guessing and Solution Behavior Based on Item Response Times in PISA

Michalis Michaelides; Militsa Ivanova, University of Cyprus

The study examines the extent of rapid guessing behavior on different types of items in the PISA Mathematics assessment, as well as the level of examinees' test-taking effort and its relationship with their test performance. We find significant variability in rapid guessers per item type and across countries.

Learning Loss During COVID-19: An Interpretation Using Vertical Scales Yu (Tracy)

Zhao; Yao Xiong, Pearson Assessments; Steven Fitzpatrick; David Shin, Pearson

This study seeks to understand the COVID-related learning loss through the lens of vertical scales. Large-scale assessment data shows that by Spring 2021, students lost about one year's learning in mathematics. In reading, younger students lost an entire year or more while older students experienced less or none learning loss.

Latent Transition Cognitive Diagnosis Model with Covariates: A Three-step Approach

Qianru Liang, The University of Hong Kong; Jimmy de la Torre, University of Hong Kong; Nancy Law, University of Hong Kong

This study proposes a bias-corrected three-step estimation approach for latent transition cognitive diagnosis models with covariates. This approach assesses changes in attribute mastery status and evaluates covariate effects on initial state and transition probabilities over time by logistic regression. The simulation study showed that this method yielded accurate parameter estimates.

037. Past Presidents Dinner (Invite Only)

NCME Annual Conference
Business Meeting
6:30 to 8:00 pm ET
Pathable: Virtual 1

038. Addressing the Data Challenges from Next-generation Assessments: Data Science Upskilling for Psychometricians (Part 1)

Training Session
1:00 to 5:00 pm ET
Pathable: Virtual 1

Digitally Based Assessments (DBAs) offer promising opportunities into insights of test takers' response process information. Yet the significantly increased volume, velocity, and variety of data pose new challenges to psychometricians for handling, analyzing, and interpreting the data to materialize their value. Data science is an emerging interdisciplinary field aimed at obtaining such insights from structured and unstructured data. Data science techniques and practices could and should be adopted into the toolkit of next generation psychometrics to help address the data challenges accompanying DBAs. This workshop is intended on providing basic data science skills and modeling strategies in the context of DBAs to help psychometricians and data analysts become better equipped to work with the increasingly big and complex data. The workshop will use Python, the dominant programming language in data science.

Presenters:
Oren Livne, Educational Testing Service
Jiangang Hao, Educational Testing Service

039. Next-Generation Cognitive Diagnosis for Small Educational Testing Settings: Innovations and Implementation (Part 1)

Training Session
1:00 to 5:00 pm ET
Pathable: Virtual 2

The training sessions provide theoretically sound and practically useful methods of cognitive diagnosis (CD) with a focus on implementations in small-scale educational settings like the classrooms. Three objectives are established. The first objective concerns the construction of the Q-matrix, including the properties of complete Q-matrices, Q-matrix validation and estimation. The second objective focuses on the development of the nonparametric classification methods for small education programs, where the number of examinees is simply too small that parameter-based estimation methods fail in analyzing the data. The third objective aims to relate CD research to the practical issue of selecting a best CD method based on the specific features of the data and the development of the nonparametric CD-CAT methods tailored to the use in small educational settings. The goal of the training sessions is to familiarize participants with recently innovations for cognitive diagnosis and to provide hands-on experiences in the R programs implementing these methods and shiny webpages. The training sessions are of interest to anyone who wishes to use or research CD in small-scale educational settings. Basic knowledge in Item Response Theory and prior exposure to R would be helpful, but are not strict requirements.

Presenters:
Chia-Yi Chiu, Rutgers, The State University
Hans Friedrich Koehn
Miguel A. Sorrel, Universidad Autónoma de Madrid
Pablo Nájera, Autonomous University of Madrid
Yu Wang, Rutgers University - GSE
Jiaxi Wang, Rutgers, The State University of New Jersey

040. Data Visualization and Analysis in the Era of COVID-19 (Part 1)

Training Session
1:00 to 5:00 pm ET
Pathable: Virtual 3

The COVID-19 pandemic disrupted student education internationally. One of the consequences of the disrupted education due to the pandemic has been cancellation, interruption and modification of the educational assessment of students. For example, in spring 2020, just after the pandemic began in the United States, all state summative testing was cancelled following the federal government issuing assessment waivers. Similarly, as student education took place remotely, interim assessment providers have altered their products to allow students to take tests while at home. These and other alterations to standard testing protocol present unique challenges to psyshometricians and data analysts who validate and use these data. This session introduces participants to numerous ways of analyzing data over the course of the pandemic that the instructors have had to develop as part of their work with state departments of education and assessment vendors. Topics include: Multiple imputation methods for addressing missing data, assessing academic impact using skip-year growth, investigating pandemic achievement gaps, and validating form comparability. Participants will install and be taught how to use customized processes from an R package (cfaTools) to analyze and validate their assessment data.

Presenters:
Damian Betebenner, National Center for The Improvement of Educational Assessment
Nathan Dadey, Center for Assessment
Adam VanIwaarden, Center for Assessment
Allie Cooperman, University of Minnesota

Virtual Sessions

041. Addressing the Data Challenges from Next-generation Assessments: Data Science Upskilling for Psychometricians (Part 2)

Training Session
1:00 to 5:00 pm ET
Pathable: Virtual 1

Digitally Based Assessments (DBAs) offer promising opportunities into insights of test takers' response process information. Yet the significantly increased volume, velocity, and variety of data pose new challenges to psychometricians for handling, analyzing, and interpreting the data to materialize their value. Data science is an emerging interdisciplinary field aimed at obtaining such insights from structured and unstructured data. Data science techniques and practices could and should be adopted into the toolkit of next generation psychometrics to help address the data challenges accompanying DBAs. This workshop is intended on providing basic data science skills and modeling strategies in the context of DBAs to help psychometricians and data analysts become better equipped to work with the increasingly big and complex data. The workshop will use Python, the dominant programming language in data science.

Presenters:
Oren Livne, Educational Testing Service
Jiangang Hao, Educational Testing Service

042. Next-Generation Cognitive Diagnosis for Small Educational Testing Settings: Innovations and Implementation (Part 2)

Training Session
1:00 to 5:00 pm ET
Pathable: Virtual 2

The training sessions provide theoretically sound and practically useful methods of cognitive diagnosis (CD) with a focus on implementations in small-scale educational settings like the classrooms. Three objectives are established. The first objective concerns the construction of the Q-matrix, including the properties of complete Q-matrices, Q-matrix validation and estimation. The second objective focuses on the development of the nonparametric classification methods for small education programs, where the number of examinees is simply too small that parameter-based estimation methods fail in analyzing the data. The third objective aims to relate CD research to the practical issue of selecting a best CD method based on the specific features of the data and the development of the nonparametric CD-CAT methods tailored to the use in small educational settings. The goal of the training sessions is to familiarize participants with recently innovations for cognitive diagnosis and to provide hands-on experiences in the R programs implementing these methods and shiny webpages. The training sessions are of interest to anyone who wishes to use or research CD in small-scale educational settings. Basic knowledge in Item Response Theory and prior exposure to R would be helpful, but are not strict requirements.

Presenters:
Chia-Yi Chiu, Rutgers, The State University
Hans Friedrich Koehn
Miguel A. Sorrel, Universidad Autónoma de Madrid
Pablo Nájera, Autonomous University of Madrid
Yu Wang, Rutgers University - GSE
Jiaxi Wang, Rutgers, The State University of New Jersey

043. Data Visualization and Analysis in the Era of COVID-19 (Part 2)

Training Session
1:00 to 5:00 pm ET
Pathable: Virtual 3

The COVID-19 pandemic disrupted student education internationally. One of the consequences of the disrupted education due to the pandemic has been cancellation, interruption and modification of the educational assessment of students. For example, in spring 2020, just after the pandemic began in the United States, all state summative testing was cancelled following the federal government issuing assessment waivers. Similarly, as student education took place remotely, interim assessment providers have altered their products to allow students to take tests while at home. These and other alterations to standard testing protocol present unique challenges to psyshometricians and data analysts who validate and use these data. This session introduces participants to numerous ways of analyzing data over the course of the pandemic that the instructors have had to develop as part of their work with state departments of education and assessment vendors. Topics include: Multiple imputation methods for addressing missing data, assessing academic impact using skip-year growth, investigating pandemic achievement gaps, and validating form comparability. Participants will install and be taught how to use customized processes from an R package (cfaTools) to analyze and validate their assessment data.

Presenters:
Damian Betebenner, National Center for the Improvement of Educational Assessment
Nathan Dadey, Center for Assessment
Adam VanIwaarden, Center for Assessment
Allie Cooperman, University of Minnesota

Virtual Sessions

044. Process Squared: Mining the processes in NAEP Process Data

Training Session
1:00 to 5:00 pm ET
Pathable: Virtual 1

The National Assessment of Educational Progress (NAEP) has transitioned to digitally based assessments (DBAs) in 2017. Availability of process data with the introduction of DBAs creates numerous possibilities for psychometricians and researchers interested in examining the detailed logs of students' interactions. Focusing on processes rather than students' final answers and responses provides an additional source of validity evidence (AERA et al., 2014), highlights different sources of success and failure on responses, and supports different assessment stages including but not limited to item development and validation of scoring rubrics. Flexible digital assessment environments such as the one NAEP uses can pose several challenges to the process analyses as the resulting processes can be messy and hard to interpret. NAEP process data consist of time-stamped records of student actions (e.g., highlighter use) as well as item specific actions (e.g., drag and drop actions). These time-stamped records form action sequences and can be used to extract meaningful information on students' response processes. This session will introduce multiple techniques from the process mining framework to be applied to NAEP process data and will provide necessary guidance on how to appropriately prepare and analyze this new data type to discover hidden patterns of response processes.

Presenters:
Emmanuel Sikali
Ruhan Circi, American Institutes for Research
Juanita Hicks, AIR
Tiago A. Caliço, American Institutes for Research

Virtual Sessions

045. Tools and Strategies for the Design and Evaluation of Interactive Dashboard Reports

Training Session
1:00 to 5:00 pm ET
Pathable: Virtual 1

Score reports are often the primary means by which score users receive information about test-takers' performance on a test. Therefore, it is critical that the information communicated in reports is iteratively evaluated to ensure that stakeholders are able to interpret and use the information in appropriate ways. More recently interest around interactive reporting systems (or dashboard reports) has been burgeoning which is apropos given the current shift towards a predominantly digital and customized world. In this workshop, we will use the iterative multistep framework (Hambleton & Zenisky, 2013; Zapata-Rivera et al., 2012) for score report design and apply this framework to discuss the various research-based methods that should be considered in the development and evaluation of dashboard reports. This training session is intended to offer practitioners the tools, strategies, and best practices they need to iteratively evaluate dashboards that are considered useful and interpretable by stakeholders in different contexts. In this session, we will focus on parents, teachers and administrators, and policy makers as three focal stakeholder groups who receive reports on a K-12 assessment. We will use various practical hands-on activities interspersed with lecture. Participants should bring their own laptops to engage in some of the practical hands-on sessions.

Presenters:
Priya Kannan, Educational Testing Service
April Zenisky, University of Massachusetts Amherst
Diego Zapata-Rivera, Educational Testing Service
Rich Feinberg, National Board of Medical Examiners

046. Applications of Keystroke Logging in Educational Research

Training Session
1:00 to 5:00 pm ET
Pathable: Virtual 2

In this half-day workshop, participants will have an opportunity to learn about and analyze a newer type of the educational data that is being progressively used in the educational research community; namely, the keystroke logs collected during the responding process on digital-based assessments. Take writing as an example, information contained in the keystroke logs goes much beyond a holistic evaluation on the final written product. From the keystroke logs, one may identify, for example, whether a writer had trouble with word retrieval, edited what was written before the submission, or experienced keyboarding difficulty. As much as the opportunities and potential applications offered by this type of log data, it also poses challenges to researchers and practitioners, including data pipeline development, construct-relevant evidence extraction, statistical treatment of such complex data, as well as making proper inferences based on the observables. Graduate students and professionals in the areas of educational measurement and writing research are invited. The format of this workshop will be a mix of capability demonstration, lecture-style presentation, hands-on data analysis, and group discussion. Some background on statistical analysis will be preferred. Sample Python and R codes will be provided. Participants should bring personal laptops with Python/Anaconda and R installed.

Presenters:
Mo Zhang, Educational Testing Service
Hongwen Guo, Educational Testing Service
Xiang Liu, Educational Testing Service

Training Sessions

047. Building Custom Interactive Dashboards with Shiny: A Tutorial with Examples

Training Session
8:00 to 4:30 pm PT
Westin San Diego Gaslamp: California Ballroom A

In many testing, commercial, and academic contexts, there is a need to summarize, visualize, and share results obtained from large data sets with various stakeholders. Dashboards allow you to structure these outputs in an accessible and interactive way, enabling end-users to navigate and explore data much easier than traditional static reports. The Shiny package in R is a free, open-source resource tool that enables users to build dashboards with a high degree of customizability and interactivity while only needing a moderate level of R programming skill. Shiny dashboards can be hosted on a local or web server, and end users do not need to know any R programming to successfully navigate them. In this session, we guide attendees through building a simple interactive dashboard in Shiny. First, we discuss the foundations of a Shiny dashboard. Second, we provide examples of different inputs within the Shiny environment. Finally, we showcase some of the advanced Shiny dashboards currently operational at different testing organizations. Attendees are encouraged to have at least a moderate level of R programming ability. More advanced R programmers will still benefit from the Shiny dashboard information, especially if they have little to no experience with building dashboards in Shiny.

Presenters:
Thai Quang Ong, National Board of Medical Examiners
Rich Feinberg, National Board of Medical Examiners
Matt Roumaya, National Board of Medical Examiners
Marcus Walker, National Commission on Certification of Physician Assistants

048. Applications of Language models in Educational Assessment

Training Session
8:00 to 4:30 pm PT
Westin San Diego Gaslamp: Plaza A

Natural Language Processing has recently experienced a renaissance with the availability of pretrained neural network-based language models. As researchers continue to push the boundaries of what language models are capable of, education stands to benefit greatly from the incorporation of neural networks in assessment, formative feedback systems, and the development of next-generation assessments. The goal of this lecture-stye training workshop is to provide a gentle introduction to neural network-based language models with a focus on their applications to educational assessment. The first part of the training session will be a theoretical introduction to the principals of machine learning, text-classification, Language models in general, and their applications in the context of educational assessment. The second part of the session will focus on getting the participant familiar with the practical use of machine learning libraries and language models in Python. Upon completion the participants should be familiar with the basic use of language models to perform short answer scoring and automated essay scoring. We finish the workshop by discussing how we gauge and monitor the quality and equity of the resulting systems.

Presenters:
Christopher Ormerod, Cambium Assessment
Amir Jafari, Cambium Assessment
Susan Lottridge, Cambium Assessment, Inc

049. Bayesian Networks in Educational Assessment (Book by Springer)

Training Session
8:00 to 4:30 pm PT
Westin San Diego Gaslamp: California Ballroom C

The Bayesian paradigm provides a convenient mathematical system for reasoning about evidence. Bayesian networks provide a graphical language for describing complex systems, and reasoning about evidence in complex models. This allows assessment designers to build assessments that have fidelity to cognitive theories and yet are mathematically tractable and can be refined with observational data. The first part of the training course will concentrate on Bayesian net basics (using Netica), while the second part will concentrate on model building and recent developments in the field. (Book is included).

Presenters:
Duanli Yan, ETS
Russell G Almond, Florida State University
Diego Zapata-Rivera, Educational Testing Service

Training Sessions

050. Using SAS for Monte Carlo Simulation Studies in Item Response Theory

Training Session
8:00 to 12:00 pm PT
Westin San Diego Gaslamp: Del Mar

Data simulation and Monte Carlo simulation studies are important skills for researchers and practitioners of educational measurement, but there are few resources on the topic. This four-hour workshop presents the basic components of Monte Carlo simulation studies (MCSS). Multiple examples will be illustrated using SAS including simulating total score distribution and item responses using the two-parameter logistic IRT, bi-factor IRT, and hierarchical IRT. Material will be applied in nature with considerable discussion of SAS simulation principles and output. The intended audience includes researchers interested in MCSS applications to measurement models as well as graduate students studying measurement. Comfort with SAS base programing and procedures will be helpful. Participants are encouraged, but not required, to bring their own laptops. The presentation format will include a mix of illustrations, discussion, and hands-on examples. As a result of participating in the workshop, attendees will: 1) Articulate the major considerations of a Monte Carlo simulation study, 2) Identify important SAS procedures and techniques for data simulation, 3) Adapt basic simulation techniques to IRT-specific examples, and 4) Extend examples to more complex models and scenarios.

Presenters:
Brian C Leventhal, James Madison University
Allison Ames Boykin, University of Arkansas

051. Analyzing NAEP/TIMSS Data with Direct Estimation in R, Theory and Practice

Training Session
8:00 to 4:30 pm PT
Westin San Diego Gaslamp: La Jolla

The course consists of two parts: (1) instructions on the theory of psychometric and sampling designs of NAEP and TIMSS assessments, and (2) the demonstration of data analysis procedures and hands-on practice, including using the plausible values and the direct estimation with item response data. Public-use NAEP and TIMSS data files will be used for demonstration and hands-on practices via the R packages EdSurvey and Dire. Both packages were developed for analyzing large-scale assessment data with complex psychometric and sampling designs. Participants will learn how to perform: • data process and manipulation • descriptive statistics • linear regression on a latent variable or plausible values The knowledge and analytic approach learned from this course can be applied to analyzing other large-scale assessment data with plausible values, including generating novel plausible values that account for additional or imputed covariates. To get the most out of the training participants should have their own computers preloaded with the latest version of the R and RStudio software to participate in the hands-on portion. This course is designed for individuals in government, universities, private sector, and nonprofit organizations who are interested in learning how to analyze large-scale assessment data through the direct estimation approach.

Presenters:
Ting Zhang, American Institutes for Research
Emmanuel Sikali
Paul Bailey, American Institutes for Research
Eric Buehler, AIR
Michael Lee, AIR

052. Applying Data Mining Methods to Detect Test Fraud

Training Session
8:00 to 12:00 pm PT
Westin San Diego Gaslamp: Sierra A

This session will provide audience with systematic training on applying various data mining models using software programs: R and/or Python. It covers the basics of these two software programs, theories of selected unsupervised and supervised learning methods, including K-Means, Gaussian Finite Mixture, Self-Organization Mapping, K-Nearest Neighbor, Random Forest, Supported Vector Machine, Neural Network with R/Python demonstration on applying them to detect test fraud. Further, the advantages and disadvantages of using each software program will be discussed. This session consists of lectures, demonstrations, and hands-on activities of running various commonly used data mining methods. It is intended for intermediate and advanced graduate students, researchers, and practitioners who are interested in learning the basics and advanced topics related to data mining methods. It is expected the audience will have some basic knowledge of R and Python programming, but not required. Attendees will bring their own laptop and download the software programs free online. It is expected that attendees will master the basics of specify various data mining models and applying these models to detect aberrantly behaved test-takers; further, they can apply the skills to their own research and datasets.

Presenters:
Kaiwen Man, University Of Alabama
Cheng Hua, University Of Alabama
Qingzhou Shi, University Of Alabama
Mehdi Rajeb, University of Alabama

Training Sessions

053. Computerized Multistage Testing: Theory and Applications (Book by Chapman and Hall)

Training Session
8:00 to 12:00 pm PT
Westin San Diego Gaslamp: Sierra B

This course provides a general overview of a computerized multistage test (MST) design and its important concepts and processes. The MST design is described, why it is needed, and how it differs from other test designs, such as linear test and computer adaptive test (CAT) designs, how it works, the methodologies involved, and its simulations.(Book is included)

Presenters:
Duanli Yan, ETS
Alina A von Davier, Duolingo
Kyung (Chris) T. Han, Graduate Management Admission

054. Using Stan for Bayesian Psychometric Modeling

Training Session
8:00 to 4:30 pm PT
Westin San Diego Gaslamp: Santa Fe

This session will provide audience with systematic training on Bayesian estimation of classic and new psychometric models using Stan. The estimation of model parameters for various psychometric models will be illustrated and demonstrated using Stan, with a particular emphasis on IRT models. Further the advantages and disadvantages of Stan comparing to traditional Bayesian software programs such as OpenBUGS and JAGS will be discussed. This session consists of lecture, demonstration, and hands-on activities of running Stan. It is intended for intermediate and advanced graduate students, researchers, and practitioners who are interested in learning the basics and advanced topics related to parameter estimation of common psychometric models using Stan. It is expected the audience will have some basic knowledge of the Bayesian theory, but not required. Attendees will bring their own laptop and download the software program free online. It is expected that attendees will master the basics of writing Stan codes in running standard and extended psychometric models; further they can develop Stan codes for new psychometric models for their own research and psychometric modeling.

Presenters:
Yong Luo
Xin Qiao, Southern Methodist University

055. Cognitive Diagnosis Modeling: A General Framework Approach and Its Implementation in R

Training Session
8:00 to 4:30 pm PT
Westin San Diego Gaslamp: Plaza B

The primary aim of the workshop is to provide participants with the necessary practical experience to use cognitive diagnosis models (CDMs) in applied settings. Moreover, it aims to highlight the theoretical underpinnings needed to ground the proper use of CDMs in practice. In this workshop, participants will be introduced to a proportional reasoning (PR) assessment that was developed from scratch using a CDM paradigm. Participants will get a number of opportunities to work with PR assessment-based data. Moreover, they will learn how to use GDINA, an R package developed by the instructors for a series of CDM analyses (e.g., model calibration, evaluation of model appropriateness at item and test levels, Q-matrix validation, differential item functioning evaluation). To ensure that participants understand the proper use of CDMs, the theoretical bases for these analyses will be discussed. The intended audience of the workshop includes anyone interested in CDMs who has some familiarity with item response theory (IRT) and R programming language. No previous knowledge of CDM is required. By the end of the session, participants are expected to have a basic understanding of the theoretical underpinnings of CDM, as well as the capability to conduct various CDM analyses using the GDINA package.

Presenters:
Jimmy de la Torre, University Of Hong Kong
Wenchao Ma, University Of Alabama

056. Non-commercial IRT-based Simulation Software: WinGen3, SimulCAT, MSTGen, and IRTEQ

Training Session
1:00 to 4:30 pm PT
Westin San Diego Gaslamp: Del Mar

This training session introduces four item response theory (IRT)-based simulation computer programs: (1) WinGen3 for generating IRT parameters and item responses, (2) SimulCAT for simulating computer adaptive testing administrations, (3) MSTGen for simulating multistage testing administrations, and (4) IRTEQ for implementing IRT equating. These software tools support various IRT models and comprehensive features with an intuitive, user-friendly interface. Through the session, attendees will have a better understanding of the importance of IRT-based simulation as well as the practical constraints and challenges of simulation-based research. The current training delivers essential psychometric knowledge and professional simulation skills, as well as passes down the practical tips to write well-defined and impactful research questions for the simulation study. The workshop is intended for junior-level practitioners and graduate students who need to conduct comprehensive data simulations in educational research. It is recommended for participants to have some background knowledge in IRT including DIF, scaling and equating, and adaptive testing, but not required. Demonstrations and hands-on practice will be conducted with proposed non-commercial software. Attendees should bring their own laptops and the most recent version of four programs installed (www.hantest.net). Instructors will send electronic training materials via email at least one week prior to the conference.

Presenters:
Hanwook Yoo, Educational Testing Service
Kyung (Chris) T. Han, Graduate Management Admission
Hyeon-Joo Oh, Educational Testing Service

057. Sequence Mining Methods on Process Data in Large-Scale Assessments

Training Session
1:00 to 4:30 pm PT
Westin San Diego Gaslamp: Sierra A

This training session introduces basic knowledge in sequence mining methods that could be used to tame complex process data in sequential format associated with timestamp and highlights advanced applications of sequence mining in analyzing process data to better support group-level (in)variance explorations in international and national assessments. Specifically, the presenters will focus on four subtopics, including (1) how to extract gram-based features from clickstream sequence, (2) how to identify similarity between any pair of sequences by computing sequence distance, (3) how to visualize sequence grouping and measure group (in)variance, and (4) how to use latent sequence models (e.g., hidden Markov model) to identify process states. During the half-day workshop, participants will be provided with an overview of process data collected from computer-based large-scale assessments, learn about various approaches to analyzing and using log data with sequence mining methods, and obtain hands-on experience working with sequential process data through examples and exercises. Intended audience are researchers, students, and practitioners with basic knowledge of process data and familiarity with R/RStudio/Python and interested in learning or applying data-driven methods to log data analysis.

Presenters:
Qiwei He, Educational Testing Service, USA
Esther Ulitzsch, IPN - Leibniz Institute for Science and Mathematics Education, Germany
Bernard Veldkamp, University of Twente, the Netherlands

058. Embedded Alignment and Standard Setting in Practice

Training Session
1:00 to 4:30 pm PT
Westin San Diego Gaslamp: Sierra B

This session will engage participants in an interactive and hands-on application of Embedded Alignment and Standard Setting (EASS) methods, including the three critical EASS processes that support assessment system coherence: 1. The alignment of test items to evidence statements articulated in achievement levels (Forte, 2017), 2. Embedded Standard Setting (ESS) analyses (Lewis & Cook, 2020), and 3. The resolution of items whose hypothesized alignments are not supported by empirical data (Lewis & Cook, 2020; Brice, 2021). We will begin the session with an introduction to EASS methods. Next, we will use a common set of test items to guide participants through an application of the three integrated EASS activities. Participants will learn: 1. why and how to align items to specific achievement levels; 2. how to use ESS software to estimate ESS cut scores and evaluate their efficacy; and 3. how to use item-level data and claim-level inconsistent item summaries to resolve ESS-inconsistent items using construct-based rationales. This session is intended for measurement professionals who would like to use alignment and standard setting methods appropriate to a principled assessment design framework. Laptops will be required to run training versions of the proprietary ESS software (EmStanS; Lewis & Lee, 2021).

Presenters:
Daniel Lewis, Creative Measurement Solutions LLC
Ellen Forte, edCount, LLC
Amanda Brice, Curriculum Associates

NCME Board of Directors Meeting #1 (Invite Only)

3:00 to 6:00 pm PT
Westin San Diego Gaslamp: Imperial

059. Educational Measurement: How We Got to Now

Coordinated Paper Session
8:00 to 9:30 am PT
Westin San Diego Gaslamp: California Ballroom A

The purpose of this symposium is to acquaint educational measurement professionals, students, and consumers of educational testing products and documentation with the historical underpinnings of current educational testing theory and practice. Its premise is that a clear understanding of the present and future of the field requires a solid understanding of its past and the trajectory by which we got to the present. The symposium consists of four papers focusing on 1) the history of norm- and criterion-referenced testing, 2) evolving notions of fairness in testing 3) a history of classical test theory, and 4) a history of item response theory. Collectively, the presentations address such diverse topics as Darwin's theory of evolution, intelligence testing, standard setting, admissions testing, accountability testing, various item response theory (IRT) models and applications, and the men and women who pioneered these fields. At the same time, the presenters will highlight the events and evolving public attitudes that shaped educational measurement from the late 19th century to the present.

Session Organizer:
Michael Brannen Bunch, Measurement Incorporated

Moderator:
Michael Brannen Bunch, Measurement Incorporated

Participants:

Norm-Referenced Testing /Criterion-Referenced Testing
Kurt F Geisinger, University Of Nebraska, Lincoln

Evolving Notions of Fairness in Testing in the U.S
Jennifer Randall, University of Massachusetts

A History of Classical Test Theory
Brian Clauser, National Board Of Medical Exam

A History of Item Response Theory
Richard Melvin Luecht

Discussant:
Robert Brennan, University Of Iowa

060. Various Ways of Using Context and Process Data in Measurement

Coordinated Paper Session
8:00 to 9:30 am PT
Westin San Diego Gaslamp: California Ballroom B

There have been studies and guidelines that have used context and process data in measurement. The value of incorporating context and process data is in enhancing the validity of the scores by controlling for or including the effect in the calculation and interpretation of test scores. In this session three different ways of using context and process data will be presented along with empirical evidence of their impact. The first way is the use of data during testing in the scoring and interpretation of the scores in an operational low-stakes interim testing program. The second way is the use of context data surrounding the activities of teachers in the development of fidelity of implementation for use in interpreting results in an instructionally embedded assessment for students with significant cognitive disabilities. The third way is the use of context data of academic preparation to estimate summative test scores in high school that show decreases in sub-group test performance gaps. Additionally, an expert discussant will present perspectives on the use of context and process data in measurement, as well as offer suggestions to enhance the methodologies reported in the three presentations.

Session Organizer:
Thanos Patelis, Teachers College at Columbia University & University of Kansas

Chair:
Jennifer Kobrin, ATLAS: University of Kansas

Participants:

Multiple Applications of Process Data to Improve Test Score Validity
Steven Wise, NWEA

A Model for Measuring Implementation Fidelity for an Instructionally Embedded Assessment System
Jennifer Kobrin, ATLAS: University of Kansas

Incorporating Academic Preparation as a Covariate in Calculating Test Scores
Thanos Patelis, Teachers College at Columbia University & University of Kansas

Discussant:
Brent Bridgeman, ETS

061. Connecting COVID-19 Policy to Practice in Indiana using State Assessment Data

Organized Discussion
8:00 to 9:30 am PT
Westin San Diego Gaslamp: California Ballroom C

Indiana, like all states, recognized that the pandemic severely disrupted its education system but grappled with the challenge of evaluating the academic impact on students as well as appropriately responding to these impacts with intentional, targeted, and evidence-based resources. To address this challenge, the Indiana General Assembly passed House Enrolled Act 1514- 2021, and Governor Eric Holcomb signed into law Public Law 211-2021. Public Law 211-2021 added Ind. Code § 20-26-5-40.6, which requires the Indiana Department of Education (IDOE) to conduct a "learning loss study" for the 2020-2021 and the 2021-2022 school years to understand the academic impact the pandemic had on Indiana students, recommend supports and resources to ameliorate these impacts, and monitor the recovery of Indiana students IDOE collected assessment data for the study from several sources including the ILEARN state summative assessment, the WIDA-ACCESS English Language Proficiency examination, and interim assessment data from multiple vendors. This session details how the IDOE used assessment data to understand the academic impact on Indiana students and ultimately connect education policy with the resources and supports necessary to help students recover.

Session Organizer:
Damian Betebenner, National Center for the Improvement of Educational Assessment

Presenters:
Maggie Paino, Indiana Department of Education
Charity Flores, Indiana Department Of Education
Adam VanIwaarden, Center for Assessment

062. Item Innovations in CAT

Paper Session
8:00 to 9:30 am PT
Westin San Diego Gaslamp: Del Mar

Participants:

A Smarter CAT: Integrating Engagement into the CAT Item Selection Algorithm
Guher Gorgun, University of Alberta; Okan Bulut, University of Alberta
In this study, we propose to integrate test-taking engagement into the CAT item selection algorithm by modifying the maximum Fisher information (MFI) method. The proposed method outperformed the MFI method, suggesting that integrating engagement into the item selection might be a feasible remedy to reduce construct-irrelevant variance in a low-stakes CAT setting.

Adaptive testing item selection strategy via deep reinforcement learning approach
Pujue Wang; Hongyun Liu, Beijing Normal University
This study proposes a CAT item selection strategy based on deep reinforcement learning. The simulation results show that the new method has higher accuracy at different test lengths and for examinees of various trait levels than the traditional strategies.

Detecting Approaches to Control Item Position Effects in CAT
Ye Ma, Amazon Web Services (AWS); Deborah Harris, University of Iowa
This study proposed four approaches to mitigate the item position effects (IPE) in CAT. Comparison across all approaches with respect to ability estimation accuracy suggest that controlling IPE at pretest level and pool level can result in better ability estimation. Implications on how to control IPE in practice are provided.

Relaxed CAT: An Extended b-Matching Method for Content Relaxed Item Selection
Kevin Cappaert, Curriculum Associates; Kristin M. Morrison, Curriculum Associates; Yu Su
The current proposal presents an extended b-matching CAT item selection method that does not require strict adherence to a content blueprint. The model is evaluated by prioritizing (not eliminating) vertically scaled items not previously administered on prior tests and content from items written to the student grade level.

Discussant:
Michael R Peabody, National Association of Boards of Pharmacy

063. Software Demonstrations

Demonstration Session
8:00 to 9:30 am PT
Westin San Diego Gaslamp: La Jolla

Participants:

Automating Item Specifications from Range ALDs to Support Item Writing

Christina Schneider, Cambium Assessment; Garron Gianopulos, NWEA; Jing Chen, NWEA

We demonstrate a proof of concept in which we automate item templates/item specifications derived from Range Achievement Level Descriptor (RALD) process documentation using Python and LaTex. Sample code and LaTeX templates will be freely available via a Google Drive along with demo materials.

Survey Weighted Mixed Models with Robust Variance Estimation in R

Paul Bailey, American Institutes for Research; Ting Zhang, American Institutes for Research

The demonstration will showcase weighted mixed model functions that developed in R packages EdSurvey and WeMix. These two interdependent packages implemented mixed models using weights at multiple levels. They are unique in allowing plausible values in the dependent variable and tailed to large-scale assessment data analysis.

VEMIRT: A Variational EM Algorithm-based Shiny App for High-dimensional IRT Applications

Chun Wang, University of Washington; Gongjun Xu, University of Michigan; Jiaying Xiao, University of Washington; Ruoyi Zhu, University of Washington; Chenchen Ma, University of Michigan

We will demonstrate a new R Shiny App called VEMIRT, which uses innovative Gaussian variational expectation-maximization (EM) methods to efficiently calibrate high-dimensional item response theory models. This demonstration will be lecture type, but audiences will be given access to the App and sample data to navigate the App features.

R Package spfa for Flexible Item Response and Response Time Modeling

Yang Liu, University of Maryland, College Park; Weimeng Wang, University Of Maryland, College Park

The R package spfa implements the expectation-maximization algorithm for fitting Liu & Wang's (2021) semiparametric factor model, which can handle various response types and has been successfully applied to analyzing assessment data with both responses and response time information. It also produces high-quality graphical display of the results.

064. Multidimensional Space

Paper Session
8:00 to 9:30 am PT
Westin San Diego Gaslamp: Santa Fe

Participants:

Relationship between measures of "21st century skills" and the content underlying them

Mark Wilson, UC Berkeley; Weeraphat Suksiri, University of California, Berkeley; Linda Morell, University Of California - Berkeley; Jonathan Francis Osborne, Stanford University; Anna MacPherson, American Museum of Natural History; Sara Dozier, California State University, Long Beach

Measuring a competency often requires items from a specific content domain—thus understanding the relationship between the competency and the content is important. We examine this relationship for a specific competency (argumentation), and the content domain that serves as the context. The investigation uses learning progressions and multidimensional Rasch models.

Investigating measurement invariance of a complex academic practice across different subject areas

Mark Wilson, UC Berkeley; Weeraphat Suksiri, University of California, Berkeley; Linda Morell, University Of California - Berkeley; Jonathan Francis Osborne, Stanford University; Anna MacPherson, American Museum of Natural History; Sara Dozier, California State University, Long Beach

Measuring a competency often requires items from a specific context—thus measurement of the skill may differ between content areas. We examine this relationship for a specific competency (argumentation) and two topics (structure of matter and ecology). We use learning progressions and multidimensional Rasch models to investigate the invariance.

Robust Estimation of Ability in Multidimensional Item Response Theory

Audrey Filonczuk, University Of Notre Dame; Maxwell Hong; Ying Cheng, University Of Notre Dame

A robust estimator for a multidimensional IRT model is proposed. Simulations reveal the effectiveness of the estimator under four types of response disturbances and various test features. We demonstrate how this robust estimation can counteract aberrancies in an aberrant educational data set containing response disturbances.

A Vector Approach to Identifying Partially Overlapping Groups in a Multidimensional Space

Jonathan Weeks, Educational Testing Service

This study uses a vector-based clustering approach with data from a measure of six foundational reading skills to identify students with the same composite score but different multidimensional profiles (loosely-integrated to highly-integrated skills). The results provide support for six-factor, two-factor, and unidimensional scores, consistent with the simple view of reading.

Discussant:

Esther Ulitzsch, IPM

065. Methodological Innovations in TIMSS and PIRLS: Robust Methods, Process Data, Artificial Intelligence

Coordinated Paper Session
8:00 to 9:30 am PT
Westin San Diego Gaslamp: Plaza

New administrative and design related innovations have advanced the accessibility and quality of international large-scale assessments, but also increased their data complexity and challenged the use of traditional statistical methods. This requires new and more rigorous statistical modeling approaches, such as more general or extended item response theory models, improved differential item functioning (DIF) analyses, and machine learning techniques. This session discusses and illustrates recent methodological innovations using empirical examples from the TIMSS and PIRLS assessments as they transition to fully-digital assessment platforms and incorporate more complex designs for sampling and test administration. The session illustrates the use of a multiple group concurrent calibration models with partial invariance assumption for increasing the cross-country comparability of contextual scales, and introduces a new approach for analyzing DIF and detecting outliers in achievement data with increasing complexity. Moreover, it demonstrates how response time data can be used to examine students' nonresponse behaviors, and how machine learning approaches can be utilized for the automatic coding of image- and text-based responses to open-ended response items.

Session Organizers:
Lale Khorramdel, Boston College
Bethany Fishbein, Boston College

Chair:
Lale Khorramdel, Boston College

Participants:

Cross-Country Invariance Modelling for Self-Report Background Scales in TIMSS 2019
Katherine Reynolds, Boston College; Lale Khorramdel, Boston College

Using robust statistics and outlier detection for analyzing DIF in PIRLS 2016
Ummugul Bezirhan, Boston College; Matthias Von Davier, Boston College

Utilizing Process Data to Examine Nonresponse Behavior in TIMSS 2019
Bethany Fishbein, Boston College; Michael Martin, Boston College; Pierre Foy, Boston College

Automated Scoring in TIMSS and PIRLS using Machine Learning and Artificial Intelligence
Matthias Von Davier, Boston College; Lillian Tyack, Boston College; Ji Yoon Jung, Boston College

Discussant:
David Kaplan, University Of Wisconsin – Madison

066. Reflections on Edmund Gordon's Pedagogical Troika: Integrating Assessment, Teaching and Learning

Coordinated Paper Session
9:45 to 11:15 am PT
Westin San Diego Gaslamp: California Ballroom A

In 2013 the eminent American psychologist, Edmund Gordon, proposed a useful conceptualization of the dynamic relationship among and between assessment, teaching and learning which he called the Pedagogical Troika. As the papers in this session will note, each of these components of the troika have had independent histories. Gordon, nevertheless, has viewed them as constituent parts of a whole cloth that viewed holistically allow greater support for the diverse characteristics of learners and their needs. Integrating learning, teaching, and assessment is not a new idea, but when focused on the affirmative development of all students serious, unresolved questions remain about psychometric and instructional design frameworks needed to motivate and engage learners and teachers in cooperative uses of instructionally relevant assessment data. The session papers will focus on the opportunities and challenges of integrating assessment, teaching and learning by describing advances in cognitively based assessments for learning (CBAL), by offering examples of how a socio-cognitive, dynamic pedagogy can advance integration, by elaborating on how "proximally generated" assessment data can support dynamic, integrated teaching and learning processes, and by calling for a unification of psychometrics and instructional design to better serve the guiding conception of the Pedagogical Troika.

Session Organizer:
Howard Everson, CUNY Graduate Center

Moderator:
Stephen G Sireci, University of Massachusetts, Amherst

Participants:

Dynamic Pedagogy: A learning-centered approach to the integration of teaching, learning and assessment at the classroom level.
Eleanor Armour-Thomas, Queens College of the City University of New York

Integrating Assessment, Teaching, and Learning: The Example of CBAL
Randy Bennett, ETS

A New Vision and Operational Frameworks for Integrating Assessment, Teaching and Learning in Diverse Learning Environments
Eva L. Baker, UCLA; Madhabi Chatterji, Teachers College, Columbia University

Unifying Instructional Design and Assessment Design to Meet the Challenges of Edmund Gordon's Pedagogical Troika
Howard Everson, CUNY Graduate Center

Discussant:
Gerunda Hughes, Howard University

067. Methods to investigate the impact of COVID-19 on assessment results

Coordinated Paper Session
9:45 to 11:15 am PT
Westin San Diego Gaslamp: California Ballroom B

The COVID-19 pandemic disrupted traditional schooling for elementary and secondary students across the world. Many schools closed in spring 2020 with little time to prepare for instructing students remotely, and most students continued to receive instruction remotely for parts of the 2020-2021 school year. These unprecedented disruptions have led to widespread concern about possible impacts on student achievement. With the resumption of statewide achievement testing, data are available to investigate impacts of the pandemic on achievement, but details about how best to analyze these data remains a challenge. The populations of students tested in 2021 likely differ from those tested in previous years due to instructional disruptions, changes in student enrollment, and disparate emphases on testing, among others. In some states certain districts tested fewer students with the state assessment while others used district assessments. Further, some students may have been tested at home while others were tested in schools; these issues have the potential to affect scores, relative to historic trends. This session will address approaches to investigating the impact of the COVID-19 pandemic on achievement scores using data from three states. The papers will highlight methods for addressing differences in the student populations and challenges to the evaluations.

Session Organizer:
Jonathan Weeks, Educational Testing Service

Participants:

Pseudo-equivalent groups approaches to examine the impact of COVID-19 on assessment outcomes
Katherine Furgol Castellano, Educational Testing Service; Jonathan Weeks, Educational Testing Service; Matthew Johnson, ETS; Daniel McCaffrey, Educational Testing Service

Hierarchical approaches to examining the impact of COVID-19 on assessment outcomes
Alexandra Stone, University of Connecticut; Katherine Furgol Castellano, Educational Testing Service; Jonathan Weeks, Educational Testing Service; Matthew Johnson, ETS; Daniel McCaffrey, Educational Testing Service

Using multiple achievement measures to understand the effects of COVID-19 on student learning
Leslie Keng, Center for Assessment; Scott Marion, National Center for the Improvement of Educational Assessment; Daniel Silver, University of Southern California Rossier School of Education

Understanding the effects of the pandemic when the state test is not available
Chris Brandt, Center for Assessment; Susan Lyons, Lyons Assessment Consulting; Lily An, Harvard Graduate School of Education; Sophie Barnes, Harvard Graduate School of Education

Discussant:
Andrew Ho, Harvard Graduate School of Education

068. Considerations in the Design of Through Year Computer Adaptive Assessments

Coordinated Paper Session
9:45 to 11:15 am PT
Westin San Diego Gaslamp: California Ballroom C

This session investigates technical questions associated with the design of a through-year adaptive assessment (a multiple administration assessment) based on the USDOE's (2017) final regulations for ESSA's Assessments under Title I, and states' desires to redesign assessment systems to support learning recovery from COVID. The final regulations permit hybrid interim-summative designs that may include items above or below a student's grade level of record, but each student's academic proficiency must be measured based on state academic content standards for the student's grade of record. How off grade information should be used for reporting purposes is not specified in the regulations; though, some helpful information may be gleaned by investigating the regulations for testing requirements for middle schooler's enrolled in advanced mathematics classes. Paper 1 will examine the historical technical issues associated with through-course assessment models. Paper 2 overviews the technical and reporting challenges that need to be considered in the design of such systems. Papers 3 and 4 show simulation evidence of how such an assessment system can function with different routing rules and item pool sizes and specifically support providing feedback to teachers on the student's present level of functioning in state standards.

Session Organizer:
Christina Schneider, Cambium Assessment

Moderator:
Paul Nichols, Planful Learning and Assessment

Participants:
The Case for an Adaptive Through Year Assessment
Garron Gianopulos, NWEA

Design Considerations and Reporting Solutions for a Multiple Administrations Adaptive Testing System
Seung W. Choi, University Of Texas At Austin; Christina Schneider, Cambium Assessment; Daniel Lewis, Creative Measurement Solutions LLC

Impact of Item Pool Size and Distribution on a TY System
Garron Gianopulos, NWEA; Jonghwan Lee, NWEA; Sangdon LIM; Luping Niu, University of Texas at Austin; Sooyong Lee, University of Texas at Austin; Seung W. Choi, University of Texas at Austin

Impact of Test Routing Rules on Theta Estimates and Achievement Level Classifications
Jonghwan Lee, NWEA; Christina Schneider, Cambium Assessment; Garron Gianopulos, NWEA; Luping Niu, University of Texas at Austin; Sooyong Lee, University of Texas at Austin; Sangdon LIM; Seung W. Choi, University of Texas at Austin

Discussant:
Richard Luecht, UNC Greensboro

069. Advancements in CAT

Paper Session
9:45 to 11:15 am PT
Westin San Diego Gaslamp: Del Mar

Participants:
Adaptive v. Non-Adaptive Test Mode Effects on Effort, Test Anxiety, and Performance
Teresa Ober, University of Notre Dame; Cheng Liu; Yikai Lu, University of Notre Dame; Ying Cheng, University of Notre Dame

We examined effects of test mode (i.e., CAT v. LFT) on effort, test anxiety, and performance (N=219; Mage=19.83). Completing the CAT mode was associated with more effortful responding but worse performance than the LFT. State test anxiety did not vary varied based on test mode.

Dealing with item calibration error in computerized adaptive testing
Aron Fink, Goethe University Frankfurt; Andreas Frey, Goethe University Frankfurt; Christoph König, Goethe University Frankfurt

Three approaches for dealing with calibration error in computerized adaptive testing (two measurement error modeling approaches, one fully Bayesian) were compared in a Monte Carlo simulation regarding the precision of the final ability estimates. Promising results were found for each approach, with the Bayesian approach outperforming the other two.

How Truncating a CAT Session Impacts Score Validity
Yeow Meng Thum, NWEA

Evidence suggests that altering an operational CAT, e.g., terminating the test early to save time, may not just reduce score precision but may change the measured construct itself. We investigate the role of time-covarying contaminants and its implication for disassociating the test-scoring model from test scaling in psychometric practice.

Minimum Sample Size Requirements for Accurate CD-CAT Classifications
Miguel A. Sorrel, Universidad Autónoma de Madrid; Pablo Nájera, Autonomous University of Madrid; Rodrigo Schames Kreitchmann, Universidad Autonoma De Madrid; Jimmy de la Torre, University Of Hong Kong; Francisco J. Abad, Universidad Autonoma De Madrid

Small-scale cognitive diagnostic assessments can provide accurate classifications despite an imprecise item parameter estimation. We explore the sample size requirements for computerized adaptive assessments, where item parameter estimates are taken as true in subsequent samples. Practical guidelines regarding test construction are provided to soften the sample size requirements.

Discussant:
Michelle Boyer, Data Recognition Corporation

070. Process Data Topics

Paper Session
9:45 to 11:15 am PT
Westin San Diego Gaslamp: La Jolla

Participants:

Examining Response Processes Using Cognitive Interviewing: Understanding the Role of Item Features in Elicited Responses

Leanne Ketterlin Geller, Southern Methodist University; Jennifer McMurrer, Southern Methodist University; Muhammad Qadeer Haider, Southern Methodist University

In this presentation, we describe research conducted with students in grades K-2 focused on complex reasoning skills in mathematics (numeric relational reasoning and spatial reasoning). As part of a larger instrument development project, 96 cognitive interviews across these constructs were conducted to evaluate students' engagement with item features. We define engagement as students' ability to attend to the instructions provided by the interviewer and their apparent comfort with the task and activities presented for each item (e.g., manipulatives). Findings highlight the task and assessment situations that impact student engagement. Importantly, we illustrate a process for analyzing cognitive interview data to illuminate students' responding behaviors for hard to assess constructs and populations.

Going Beyond Actions: A Cognition-Centered Approach to Interpreting Pauses from Process Data

Burcu Arslan, Educational Testing Service Global B.V.; Caitlin Tenison, Educational Testing Service; Bridgid Finn, Educational Testing Service

Pauses in process data can provide additional information about student knowledge, skills, and abilities. The goal of this study is to demonstrate: (a) interpretation of pauses in a cognitively valid way, (b) application of a hybrid approach to modeling pauses that combines a top-down theory-driven approach with bottom-up data-driven approaches.

Incorporating Pauses in Process Data Modeling with Heterogeneous Hidden Markov Models

Caitlin Tenison, Educational Testing Service; Burcu Arslan, Educational Testing Service Global B.V.

In the current study, we present heterogenous hidden Markov models as a promising method for modeling both the context and timing of responses. This work has implications for how incorporate response time into process data models of complex interactive computer tasks and reflect on the cognitive processes driving student behavior.

Textual Data as Process Data: A New Scoring Procedure for Mixed-Format Assessments

Jordan M. Wheeler, University of Georgia; Shiyu Wang; Yanyan Tan; Allan Cohen, University of Georgia

Mixed-format assessments are commonly used for achievement testing. In this study, we propose a scoring procedure for mixed-format assessments that incorporates information extracted from the textual data to constructed-response items. The results of the study shows that the proposed scoring procedure yields better ability estimates than traditional IRT scoring procedures.

Using timing information to understand international achievement performance in TIMSS 2019

Yuan-Ling Liaw

With the development of digital test formats, process data and log files provide the potential of revolutionary innovative information and insight on individual problem-solving processes as well as test taking behavior and engagement. We use TIMSS data to investigate whether process data can improve the scoring process and performance measure.

Discussant:

Ruhan Circi, American Institutes For Research

071. Assessing Personal Skills and Qualities for High-Stakes Higher Education Admissions

Coordinated Paper Session
9:45 to 11:15 am PT
Westin San Diego Gaslamp: Santa Fe

The purpose of this coordinated paper session is to review the research and technical challenges associated with building and evaluating the Personal Skills and Qualities (PSQ®) assessment, a multidimensional forced-choice self-evaluation currently being pilot tested in 15 universities for use in higher education admissions. The first presentation discusses the approach taken to identify the key dimensions assessed based on a set of interviews with faculty members, a comprehensive review of dimensions from commercial inventories, and a secondary analysis of data from the International Personality Item Pool. The presentation also summarizes predictive validity findings. The second presentation discusses using mixed integer programming to create statement blocks (pairs and triples) and two parallel forms based on a Likert response calibration. The third presentation summarizes results from a comparison of two scoring models—Thurstonian IRT and multi-unidimensional pairwise preference—on model fit and reliability for both pairs and triples. The final presentation outlines the development of a scoring engine which uses numerical optimization and different approaches to choose starting values (and avoid local minima) to estimate theta scores from the set of responses and is capable of mixing responses from pairs and triples.

Session Organizer:

Patrick Charles Kyllonen, ETS

Participants:

Identification of Key Personal Skills and Qualities for Higher Education

Patrick Charles Kyllonen, ETS

Automated Block and Form Assembly for a Multidimensional Forced-Choice Assessment
Qi Diao, ETS

Pairs versus Triples, Thurstonian versus MUPP Scoring on Model Fit and Reliability
Sean Joo, University of Kansas

Real Time Scoring with Numerical Optimization
Oren Livne, Educational Testing Service

Discussant:
Paul De Boeck, OSU

072. Medical Education and Testing

Paper Session
9:45 to 11:15 am PT
Westin San Diego Gaslamp: Plaza

Participants:

Utilizing Linear Logistic Test Models to Explore Medical Certification Examination Item Characteristics
Emily Karen Toutkoushian; Huaping Sun, American Board of Anethesiolog
This study uses linear logistic test models to investigate whether and how various item characteristics are associated with item difficulty in two multiple-choice subspecialty certification exams. The results suggest several characteristics that are significant and meaningful predictors of item difficulty. This study has implications for assessment and item design.

Multilevel Item Structure Modeling on Automatically Generated Items - A Bayesian Approach
Yan Yan, Georgia Tech; Andrew Dallas, National Commission on Certification of Physician Assistants
This study applied several IRT models with additive multilevel item structure (AMIS) to a medical exam data for automatically generated items. The within and between item family (shell) variation of item difficulties were examined, potential predictors were investigated, and the most appropriate model based on the MCMC estimation was selected.

Literature Review of Formative Assessment in Medical Education
Dukjae Lee, University of Massachusetts Amherst; Francis O'Donnell, National Board of Medical Examiners; Amanda Clauser, National Board of Medical Examiners
This literature review explored how formative assessments have been used in medical education and other professional education contexts over the last two decades. Studies were gathered through database searches and snowballing. Topics include theoretical frameworks, item formats, delivery modes, scoring, score reporting, and validity evidence. Measurement challenges are also addressed.

A comparison study of linking methods for multiple groups
Seongeun Kim, University of North Carolina at Greensboro; Fen Fan, NCCPA
This study compared linking methods for multiple-group design; methods include a measurement alignment method, generalized version of moment methods, and characteristic curve methods, and concurrent calibration. The relative performance of the methods will be examined by a simulation study and empirical data analysis from a medical licensure exam.

Exploring Response Process Validity Evidence for a Medical Licensing Examination
Ni Bei, 1; Ravi Pandian, NBME; Monica Cuddy, NBME
This study examines the associations among response process testing features (highlight, strikeout, annotate, and lab values) and item-level performance on a high-stakes medical licensing examination. For a sample of 312,888 items, logistic regression was used to estimate these relationships. Results suggest that some features may be more useful than others.

Discussant:
Pooja Shivraj, American Board of OBGYN

073. Advancing and Assessing Civic Learning to Promote Equity

Coordinated Paper Session
11:30 to 1:00 pm PT
Westin San Diego Gaslamp: California Ballroom A

Recent events, including contentious arguments over COVID-19 safety protocols and schools' approaches to addressing systemic racism, have generated considerable interest in improving civic learning opportunities in America's schools. "Civic learning" can be defined broadly to encompass multiple competencies (e.g., digital information literacy, ability to engage in constructive argumentation, understanding of the responsibilities of citizenship). Civic-learning opportunities are distributed inequitably, and recent NAEP results indicate disparities in civic-learning outcomes. To address these inequities, educators and policymakers will need high-quality assessments for monitoring civic learning opportunities and outcomes both on a large scale and at the classroom level. This session brings together several research and development efforts that aim to address this need. The first presentation provides evidence of inequities in learning opportunities and presents a framework for monitoring equity in civic learning. The second presentation describes results from a study of a large school district's efforts to monitor civic learning for the purpose of continuous improvement. The third presentation discusses the integration of civic-learning measures into science classroom assessments. In the final presentation, authors present an example of a complex task to support formative assessment in civics. A discussant will share the perspectives of a state education agency leader.

Session Organizer:
Laura Hamilton, ETS

Participants:

Indicators for Monitoring Equity in Civic Learning
Laura Hamilton, ETS; Julia Kaufman, RAND Corporation

District-Wide Assessment to Support Civics Reform: Chicago's Model
Erica Hodgin, University of California Riverside; Joseph Kahne, University of California Riverside

Transforming Science Learning through Collaborative Argumentation on Civics Issues
Lei Liu, Educational Testing Service; Dante Cisterna-Alburquerque, ETS; Yi Song, Educational Testing Service

Formative Assessment Task to Support Civics Teaching and Learning
Gregory Vafis, Educational Testing Service; E. Caroline Wylie, Educational Testing Service

Discussant:
Tamara Heck, Michigan State Department of Education

074. Best Practices in Evaluating Computer Scoring of Constructed Responses for Educational Measurement

Organized Discussion
11:30 to 1:00 pm PT
Westin San Diego Gaslamp: California Ballroom B

Use of computers to score performances from standardized evaluations, such as using artificial intelligence (AI) and natural language processing (NLP) to rate written or spoken responses to standardized test items, is growing rapidly in popularity. Widespread concern and skepticism, however, remain about the ability of such automated scores to support the claims of the items and the tests and whether scores are fair and comparable across all test takers. The existing standards and guidelines cite the need for evidence on the reliability, fairness, and validity of scores that rely on automated scoring, but provide few specifics about that evidence. The goal of this panel discussion is to use a moderated conversation to elicit what can constitute useful evidence about automated scores and how to interpret the results of statistical analyses. This includes thresholds and rules of thumb for decision making. The discussion will start with a review of proposed best practices for evaluating automated scores from ETS's Best Practices in Constructed-Response Scoring (Best Practices). Panelists, who are experts in validity, fairness, and operational automated scoring, will respond to these proposed practices and suggest alternatives. This conversation will be followed by a moderated discussion with the panelists and the audience.

Session Organizer:
Daniel McCaffrey, Educational Testing Service

Moderator:
Kadriye Ercikan, Educational Testing Service

Presenters:
Steve Ferrara, Cognia
Suzanne Lane, University of Pittsburgh
Susan Lottridge, Cambium Assessment, Inc
COREY PALERMO, Measurement Incorporated

075. Test Optional Policy: The Current Trends and Impact on Applications and Admissions

Coordinated Paper Session
11:30 to 1:00 pm PT
Westin San Diego Gaslamp: California Ballroom C

The trend of going test-optional has accelerated exponentially during the Covid-related pandemic. As this trend is continuing, so is the debate around the scientific evidence in favor of test-required or test-optional practices. More research and investigations are needed to better understand how the admissions practices are conducted under test-optional policies and what are these practices' impact at the institution level and at the applicant level (e.g., application behavior). In this session, we will address related questions from multiple angles. The first paper will explore the relationship between an applicant's essays, test scores, and socioeconomic background using a data set based on 60,000 applicants. The second paper will discuss the current status of test-optional policy by analyzing 521 institutions' admissions webpages. In the third presentation, we will share our findings of a new approach involving contextualized measures to inform test-optional holistic admissions practices and policies in a Midwestern state. The fourth paper will explore the impact of test-optional policy on international students' application behaviors using years of TOEFL iBT score records. We expect the approaches and findings shared would inform further research and scientific discussions around evidence needed for a valid and fair admissions process for U.S. higher education.

Session Organizer:
Guangming Ling, Educational Testing Service

Chair:
Ou Lydia Liu, ETS

Participants:

The Interplay Between Applications Essay, Income, and SAT for College Applications
A J ALVERO, Stanford University; Ben Domingue, Stanford University; Anthony Antonio, Stanford University; Ben Gebre-Medhin, mount Holyoke College; Sonia Giebel, Stanford University; Alvin Pearman, Stanford University; Mitchell Stevens, Stanford University

The Current Status of Test-Optional Policy: An Exploration of Admissions Webpages sugene
Cho-Baker, ETS; Guangming Ling, Educational Testing Service; Teresa King, Educational Testing Service; Matthew McDevitt, ETS; Jennifer Bochenek, ETS

Using Contextualized Measures to Inform Test-Optional Holistic Admissions Practices and Policies
Michael Bastedo, University of Michigan; Mark Umbricht, University of Michigan; Emma Bausch, University of Michigan; Bo-Kyung Byun, University of Michigan

An Exploration of Test Optional Policy's Impact on International Students' Applications
Emma M. Klugman, Harvard Graduate School of Education; Guangming Ling, Educational Testing Service; Caitlin Tenison, Educational Testing Service

Discussant:
Andrew Ho, Harvard Graduate School of Education

076. Digital-transitions, Device-effects and Disadvantage in Large-Scale Assessments

Coordinated Paper Session
11:30 to 1:00 pm PT
Westin San Diego Gaslamp: Del Mar

This coordinated paper session examines whether the continual technological transitions taking place in NAEP and international assessments may be disadvantaging certain student groups. Although NAEP and international assessments have both transitioned from paper-based to digitally-based, technology continues to evolve and new technologies could threaten the validity and fairness of assessments. In the proposed session, four presentations examine theoretically and empirically whether marginalized student groups are disadvantaged in e-Assessments. The first paper provides a framework for analyzing device effects. The three papers that follow examine issues of fairness and equity in e-Assessment through an analysis of data from the NAEP 2019 Computer Access and Familiarity Study, PIRLS and ePIRLS 2016, and TIMSS 2019. Overall, the three empirical studies showed complex and surprising relationships. Results from the NAEP 2019 data showed students who use digital devices more at school had lower science achievement than their peers; PIRLS results from the United States showed students from disadvantaged backgrounds did better on ePIRLS than PIRLS; and TIMSS cross-country results showed that in most countries/grade levels analyzed, the digital device type (tablet v. computer) used in assessment administration had no significant relationship with student performance.

Session Organizer:
Martin Hooper, American Institutes for Research

Chair:
Dirk Hastedt, International Association for the Evaluation of Educational Achievement

Participants:

Framework for Considering Device and Interface Features That May Affect Student Performance
Ellen Strain-Seymour, Pearson; Walter Way, College Board

Examining Digital Technology Access and Familiarity on NAEP 2019
Bitnara Jasmine Park, American Institutes for Research; George William Bohrnstedt, AIR; Davis Cousar, AIR; Martin Hooper, American Institutes for Research; Brittany Boyd, American Institutes for Research; Sami Kitmitto, AIR

Analyzing PIRLS 2016 to Explain Differences in Online and Offline Reading Scores
Elena Forzani, Boston University; Martin Hooper, American Institutes for Research

Does Device-Type Matter? Analysis of Tablet and PC Use on TIMSS
Martin Hooper, American Institutes for Research; Yifan Bai, American Institutes for Research; Fusun Sahin, American Institutes For Research

Discussant:
Julian Fraillon, Australian Council for Education

077. Test Taking Behavior/Engagement Models

Paper Session
11:30 to 1:00 pm PT
Westin San Diego Gaslamp: La Jolla

Participants:

An Illustration of the Solution Behavior IRTree Model

Brian C Leventhal, James Madison University; Dena Pastor, James Madison University; Paulius Satkus, James Madison University

Low stakes test performance commonly reflects examinee ability and effort. Examinees exhibiting low effort may be identified as using non-solution behavior through rapid response tendencies. In this study, we introduce the SBIRTree to parse out non-solution behavior from examinee ability, done by explicitly modeling an examinee's hypothesized response process.

Evaluating Rapid Guessing Response Patterns on Multistage Assessment: A Simulation Study

Samuel Dale Ihlenfeldt, University of Minnesota; Jiayi Deng, University of Minnesota

Noneffortful rapid guessing (RG) on multistage assessment undermines the validity of inferences made from ability estimates. Multiple response patterns were simulated for simulees of varying ability to explore estimation accuracy based on the location of RG. An effort-moderated IRT model is explored as a potential solution. Implications are discussed.

Rapid Responses and Item-Position Effects in Computer-Based Tests: A Joint Perspective

Marlit Lindner, Ipn Kiel, Germany; Esther Ulitzsch; Gabriel Nagy, Leibniz Institute for Science

We examined two potential indicators of unmotivated test-taking behavior, rapid responses and item position effects (i.e., performance decline across testing), in a computer-based, randomized test. Rapid responses were an important source for identifying low examinee effort, but could not fully account for the observed decline in performance across time.

Test-Taking Engagement and Test Performance: Designing a Predictive Model to Improve Performance

Seyma N. Yildirim-Erbasli, University of Alberta; Guher Gorgun, University of Alberta

This study investigates the relationship between student test-taking engagement and performance and compares four commonly used classification algorithms to design a predictive model of performance. Results suggest that a proactive model can be built to predict student performance based on their effort to intervene for improving students' engagement and thus performance.

Are They Trying? Motivation in State Census Testing with a College Admissions Exam

Jeffrey Steedle, ACT, Inc.

This study compared scores and apparent motivation before and after introducing statewide administration of a college admissions test. Scores and motivation both decreased, but motivation decreases were consistent with expectations based on motivated testers. Mean score differences between race/ethnicity groups sometimes increased and sometimes decreased.

Discussant:

Steven Wise, NWEA

078. Item Detection and Design

Paper Session
11:30 to 1:00 pm PT
Westin San Diego Gaslamp: Santa Fe

Participants:

Enemy Item Detection for Different Item Types

Yanyan Fu, GMAC; Kyung (Chris) Han, Graduate Management Admission Council

Several similarity metrics were extracted using natural language processing to predict the enemy relationship among test item pairs. The preliminary results showed that the predictive accuracy was effectively high for the sentence correction (.96) and critical reasoning (.90) item types but low for the reading-comprehension item type.

Enemy Item Detection Using Word Embedding

Xueming Li, NWEA; Ann Hu, NWEA; Gina Wilmurth, NWEA

Enemy items undermine test validity and reliability. They are identified by subject matter experts (SMEs) during item review through a manual and time-consuming process. This study presents and evaluates a method for detecting enemy items that uses word embedding and supervised learning. The results demonstrate strong model performance.

Two New Models for Item Preknowledge

Kylie Gorney, University of Wisconsin-Madison; James Wollack, University of Wisconsin

To evaluate preknowledge detection methods, researchers often conduct simulation studies in which they use models to generate the data. In this paper, we propose two new models that allow the impact of preknowledge to vary across persons and items to better represent situations that are encountered in practice.

Using a Data-driven Approach to Guide Item Development in Medical Examinations

Xia Mao, NBOME

The study investigates an approach to facilitate item development in medical examinations by analyzing item response data and scoring data for short answer questions. Real data from a medical licensure examination are used to conduct the analysis and content experts' feedback is included to refine the item writing guidelines.

Discussant:

Anna Topczewski, Law School Admission Council

079. Growth/Longitudinal Methods

Paper Session
11:30 to 1:00 pm PT
Westin San Diego Gaslamp: Plaza

Participants:

A Survey of Modern Approaches for Modeling Longitudinal Assessment

Mingqi Hu, University of Illinois, Urbana; Zhongtian Lin, Cambium Assessment, Inc

Longitudinal assessments track students' stepwise growth is gaining attraction now more than ever. Three longitudinal IRT models (L-UIRTM, LGC-IRTM, Dynamic IRTM) and one longitudinal CDM (Long-DINA) were compared, and their performance was studied with simulated data to provide theoretical and practical disentanglement when choosing among the models.

An Empirical Study of Student Growth Measure and Longitudinal Growth Projection

Catherine Xueying Francis, Houghton Mifflin Harcourt; John Denbleyker, Houghton Mifflin Harcourt; Alex Brodersen, Houghton Mifflin Harcourt

The study proposes a student growth model and a method of projecting students' growth longitudinally. The growth model measures against a criterion-referenced "targeted growth". The growth projection forecasts the amount of growth needed to achieve a certain performance level. Empirical data were analyzed for growth measure and projection accuracy.

Detecting Learning Progressions with Longitudinal Diagnostic Classification Models

Matthew James Madison, University of Georgia; Meredith Langi, NWEA; Yon Soo Suh, UCLA

This study proposes a model combining a longitudinal DCM and the hierarchical DCM. This fusion of models allows for the examination of longitudinal attribute hierarchies, which can be considered types of learning progressions. Specifically, we examine inferential methods for the detection of full- and partial-sample learning progressions.

Matthew Effects in the Measurement of Growth: Reality or Methodological Artifact?

Xiangyi Liao; Daniel Bolt, University of Wisconsin, Madison

A Matthew effect refers to a tendency to see a positive correlation between baseline proficiency and growth and is frequently of interest in assessment of proficiencies like reading comprehension. We illustrate how such correlations are highly sensitive to item characteristic curve (ICC) asymmetry and may frequently emerge as measurement artifacts.

Measuring Growth Using Learning Progressions: Measurement Accuracy When Adaptive Testing Is Used

Duy N. Pham, Educational Testing Service; Yong Luo, ETS; Longjuan Liang, ETS

This study examined the accuracy of a growth measure based on learning levels of a learning progression. Simulees and item banks were generated under the Rasch model framework; then the simulees were run through an adaptive assessment. Results indicated that 54 to 73 percent of true growth could be recovered.

Discussant:

Jeff M. Allen, ACT

080. Addressing Differential Educational Outcomes for Marginalized Populations in the Era of COVID-19

11:30 to 1:00 pm PT
Manchester Grand Hyatt: Seaport Ballroom E

(CODIT Session) As education researchers, one of our greatest strengths is the ability to influence how education research is conducted and what evidence is shared to increase awareness, promote empowerment, and stimulate social change. Therefore, AERA Division D Equity and Inclusion Committee, in collaboration with NCME Diversity Issues in Testing Committee, has organized a panel of speakers from a range of interdisciplinary fields and social justice interests who will examine from diverse methodological and measurement perspectives the differential impact of COVID-19 on educational outcomes for marginalized populations. Also addressed are changes to research methods that are more grounded in the lived experience and ways of knowing of our communities as we continue our journey towards "cultivating equitable educational systems for the 21st century".

Chairs:

Venessa Manna, Educational Testing Service
Raman Grover, BC Ministry of Education

Presenters:

Alexandra E Pavlakis, Southern Methodist University
Margarita Olivera Aguilar, Educational Testing Service
Jay Schyler Raadt, Marine Corps University
Pohai Kukea Shultz, University of Hawaii at Manoa
Edna Tan, University of North Carolina at Greensboro
Terran Brown, New Meridian

081. Philosophical Realism in Educational Measurement: Is There a There There? And Why Does It Matter?

Organized Discussion
1:15 to 2:15 pm PT
Westin San Diego Gaslamp: California Ballroom A

In the context of educational and psychological measurement, what is actually measured? What sorts of entities are the targets of measurement (sometimes referred to as "constructs", "attributes", "properties", or "latent variables")? In particular, are they real, and if so, in what sense? And if not, what does this imply about the inferences and actions taken on the basis of tests -- which are themselves very much real? In this talk (which, broadly, might be seen as reacting to the work of scholars such as Joel Michell, Denny Borsboom, and Bob Mislevy, as well as philosophers such as Thomas Teo), I will attempt to argue for a form of philosophical realism about the targets of educational and psychological measurement (referred to here as "psychosocial properties") with an eye towards why and how such a stance has practical consequences for how tests are constructed and validated and how their results are interpreted. In particular, I will argue that it makes sense both theoretically and pragmatically to consider psychosocial properties to be real entities, existing in the same spatiotemporally structured world as everything else, rather than as fictional or metaphorical entities that exist only in the minds of researchers and testing professionals.

Presenter:

Andrew Maul, University of California, Santa Barbara

Discussant:

Robert J. Mislevy

082. Research Blitz

Research Blitz Session
1:15 to 2:15 pm PT
Westin San Diego Gaslamp: California Ballroom B

Chair:

Peter Halpin, UNC-CHAPEL HIL

Participants:

Complexity of item response theory models with nonparametric density estimators

Yun-Kyung Kim, University of California – Los Angeles; Li Cai, UCLA

This study illustrates and verifies two approaches to considering model complexity when evaluating model fit of IRT model with a nonparametric density estimator: we can penalize the model fit for the number of parameters (i.e., parametric complexity) or for the stochastic complexity characterized by information matrix (i.e., structural complexity).

Computer-based Testing with Automated Feedback: Effects on Metacognition and Performance

Ute Mertens, Leibniz Institute for Science and Mathematics Education; Marlit Lindner, Ipn Kiel, Germany

In this experiment, we vary automated task feedback in a computer-based science test to assess the effects of three different feedback types on recall performance, error correction, and metacognitive ratings. The study is constructed to provide empirical guidance regarding the implementation of feedback in different computer-based assessment contexts.

Linking Methods for the Multi-Unidimensional Pairwise Preference (MUPP) IRT Model

Lavanya Shravan Kumar, University of South Florida; Naidan Tu, University of South Florida; Sean Joo, University of Kansas; Stephen Stark, University of South Florida

Multidimensional forced-choice (MFC) measures are gaining prominence in noncognitive assessment. Yet there has been little research on linking methods that support developing item banks and differential item functioning analysis. This research examines Haebara, Stocking-Lord, mean-mean, and mean-sigma linking with MFC tests based on the Multi-Unidimensional Pairwise Preference model.

Examining Examiner Bias Using Cross-Classified Multilevel Model

Nai-En Tang; Chia-Lin Tsai; Daniel Edi, University of Northern Colorado; Igor Himelfarb

Examiner bias (Fuchs & Fuchs, 1986) may influence the examinee's scores in a clinical performance-based exam (Guraya et al., 2010). Cross-classified multilevel models (Goldstein, 1994) were applied to a 10-station exam rated by 160 examiners for 677 examinees. The results suggested variability among examiners and examiner effect should be controlled.

For Improving Subscore Estimates in a Short-Length Computerized Adaptive Test

Unhee Ju, Riverside Insights; JongPil Kim, Riverside Insights

Although the measurement quality of overall score for a short-length CAT is adequate, the quality of subscores might not be sufficiently high for reporting, mainly due to insufficient number of items per domain. This study investigates factors to improve the measurement quality of subscore estimates in a short-length CAT.

083. Applications in Linking & Equating

Paper Session
1:15 to 2:15 pm PT
Westin San Diego Gaslamp: California Ballroom C

Participants:

A Semiparametric Model for Equating Oral Reading Fluency Scores Cornelis Potgieter, Texas

Christian University; Akhito Kamata, Southern Methodist University

This study develops a semiparametric model for equating oral reading fluency scores. The conditional number of words read correct model specifies a general mean-variance structure, allowing for over- or under-dispersed counts. Person- and passage-specific Gaussian latent factor components also allow for different correlation structures within and between reading passages.

Evaluating Equating Methods for Varying Levels of Form Difference

Ting Sun, Northwestern University; Stella Kim, University of North Carolina at Charlotte

The study aims to examine the effect of the magnitude of a form-difficulty difference on equating results. It has been found that mean or linear equating performs better with no or small difficulty difference, whereas equipercentile equating functions well when the difficulty difference is medium or large.

Evaluating Performance of Model-Based Approach for Equating Oral Reading Fluency Scores

Yusuf Kara, Southern Methodist University; Akihito Kamata, Southern Methodist University; Cornelis Potgieter, Texas Christian University; Joseph F. T. Nese, University of Oregon

This study evaluates the performance of model-based equating procedures for oral reading fluency (ORF) scores estimated by a latent measurement model. Simulation results showed that model-based ORF scores should be preferred over observed words correct per minute measures for passages that have been equated by the common-item non-equivalent group design.

Discussant:

Jon S. Twing, Pearson Assessments and Qualifications

084. Advances in Test Design/Assessment Design

Paper Session
1:15 to 2:15 pm PT
Westin San Diego Gaslamp: Del Mar

Participants:

Corrective Feedback in Computerized Assessment: Does Feedback Message Complexity Matter?

Livia Kuklick, Ipn Kiel, Germany; Samuel Greiff, University of Luxembourg; Marlit Annalena Lindner, IPN Kiel

Computer-based assessment facilitates the implementation of immediate feedback. In this experiment, we investigate the potentials and limitations of different feedback strategies for computerized low-stakes assessment. We systematically vary the complexity of feedback messages following incorrect responses and analyze the effects of error clarification complexity on students' metacognition, cognition, and motivation.

Diagnosing Student Mastery of Standards: Impact of Varying Item Response Modeling Approaches

Susan Embretson, Georgia Institute of Technology

Diagnosing students' mastery of previous grade-level standards is useful for determining needed remedial instruction. The current study applies varying IRT models to a year-end achievement test to the diagnose mastery. It was found that IRT mixture models provide precise and reliable diagnosis while accommodating varying student mastery patterns.

Discussant:

Whitney Coggeshall, American Board of Internal Medicine

085. Security, Proctoring and More

Paper Session
1:15 to 2:15 pm PT
Westin San Diego Gaslamp: La Jolla

Participants:

Applying the CUSUM Method to Detect Aberrant Responses in Credentialing Testing

Siyu Wan, UMASS Amherst; Irina Grabovsky, 3750 Market Street; Lisa Keller, University Of Massachusetts

The cumulative sum (CUSUM) is an established statistical process control method using a person-fit statistic to detect aberrant responses for educational assessments. This study applied the CUSUM method to a credentialing examination. It flagged and distinguished different types of misfitting response behaviors. The practical implication of the CUSUM was discussed.

A Randomized Controlled Study Comparing Remote Proctoring and Test Center Administrations

Hao Song, NBCRNA; Timothy Muckle, NBCRNA

Participants of a moderate-stakes continued certification assessment were controlled on major background characteristics, and randomly assigned to live remote proctoring and test center administrations. Psychometric analysis showed comparable test performances. Surveys and proctor notes reported that takers in the remote proctoring mode experienced more technical issues, but less testing anxiety.

An Empirical Study of Quantitative Methods for Detecting Cheating

Jiaying Xiao, University of Washington; Michael R Peabody, National Association of Boards of Pharmacy

This empirical study was designed to apply several four quantitative methods to exam results data and compare their ability to identify known cheaters and exposed items. Preliminary results demonstrated some implications and limitations for current psychometric methods. Further analysis to improve the method performance is discussed.

Discussant:

***John Fremer*, Caveon Test Security**

086. DIF Approaches in CAT

Paper Session
1:15 to 2:15 pm PT
Westin San Diego Gaslamp: Santa Fe

Participants:

Detecting Differential Item Functioning in CAT using IRT Residual DIF approach

Hwanggyu Lim, Graduate Management Admission Council; Edison M. Choe, Renaissance; Kyung (Chris) Han, Graduate Management Admission Council

This study investigates the performance of the IRT residual DIF (RDIF) detection approach (Lim et al., 2021) for assessing uniform DIF in CAT. A preliminary simulation study demonstrates its potential as a powerful and practical method with robust DIF detection in CAT, even with relatively small sample sizes.

Differential Item Functioning Detection for Continuous Response CAT

Ruoyi Zhu, University of Washington; Chun Wang, University of Washington

We propose two DIF detection methods---a modified computerized adaptive testing SIBTEST and a Rasch regularization method--- for continuous responses, severely sparse CAT. Severe sparsity arises when large number of items are automatically generated. Simulation studies and real data analyses using data from Duolingo English Test are conducted.

Impact of matching criterion for DIF detection in CAT assessments

Jungnam Kim; Hongwook Suh, Nebraska Department of Education; Nisha Padminiamma, NWEA; Melinda Montgomery, Pearson

This study investigates the effect of matching criterion and detection methods for assessing DIF in statewide CAT assessments. The analyses included various student groups of interests and test data with different assessment designs. Analyses results and further discussion will provide practical implications in DIF analyses in CAT assessment settings.

Discussant:

Yong He, Measurement Incorporated

087. Survey Research

Paper Session
1:15 to 2:15 pm PT
Westin San Diego Gaslamp: Plaza

Participants:

(Self)-Reporting for Duty: Comparing Measurement Paradigms in Developing Liberal Arts Inclination scale
Gabe Avakian Orona, UC Irvine

Self-reports are commonly used to measure a slew of educational outcomes and processes. However, recent work has shown that common psychometric applications cannot be relied upon to make accurate inferences regarding measurement quality. The current paper illuminates further issues of self-reports via an empirical comparison of popular measurement paradigms.

Overclaiming Technique - Method to Increase Self-Reports' Measurement Quality in Educational Research?
Marek Muszyński, IFiS PAN; Tomasz Żółtak, IFiS PAN; Artur Pokropek, IFiS PAN

Questionnaires (self-reports) are an important source of data in many types of educational research, from large-scale international assessments to smaller projects. However, its measurement qualities are often compromised by response biases that decrease its predictive validity and international comparability. Overclaiming technique is tested as a remedy to these problems.

Discussant:
Susan Davis-Becker, ACS Ventures, LLC

088. What Does Philosophy and History Have to Contribute to Educational Measurement?

Coordinated Paper Session
2:30 to 4:00 pm PT
Westin San Diego Gaslamp: California Ballroom A

Some might argue that, in a field that has increasingly focused on technological solutions to problems in large-scale testing, matters of philosophy and history are merely academic, and should not take up precious conference time. Others might counter that pursuing technological solutions without a firm basis in the thought and logic of the application area is an unwise strategy. Some might comment that, while the development and application of technological solutions to large-scale testing challenges has been the bread and butter of many NCME members, at a time when we are moving into a wider scope for our measurement expertise, taking on challenges involved in classroom testing and in online testing contexts (where process measurements are important), we should be re-assessing our foundations to make sure that they are sound and sufficient to address these new frontiers. Others, being old-fashioned academics, might be simply embarrassed by this apparent failure to maintain appropriate academic standards. In response to such considerations, this session will consist of summaries of four recent efforts to update and solidify the foundations of the theory and practice of educational measurement.

Session Organizer:
Mark Wilson, UC Berkeley

Moderator:
Brian Clauser, National Board of Medical Examiners

Participants:

Historical and Conceptual Foundations of Measurement in the Human Sciences
Derek Briggs, University of Colorado

A Pragmatic Perspective of Measurement
David Torres Irribarra, Pontifical Catholic University of Chile

Measurement Across the Sciences: Developing a Shared Concept System for Measurement
Mark Wilson, UC Berkeley

On the Nature of Measurement
Joshua McGrane, University of Oxford

089. Methods and Results in Monitoring Hybrid Automated/Hand-scoring of Essays

Coordinated Paper Session
2:30 to 4:00 pm PT
Westin San Diego Gaslamp: California Ballroom B

Most automated scoring engines are used alongside human raters in live scoring events as a way to produce aggregate scores from engines and humans, monitor the human or engine scoring via second reads by the other, or to route engine-scored unusual responses for human rating. In some cases, all responses receive scores from a human rater and an engine; in others, only a portion of responses (e.g., 10-25%) receive a human rating. While the monitoring of human rater performance is well-studied, monitoring of hybrid scoring has seen less attention. In this symposium we hope to provide some insight into best practice in designing systems for hybrid scoring, present studies that illustrate hybrid scoring performance in two operational settings, and end with a discussion drawing on the discussant's many years of operational hybrid and automated scoring experience. Our hope in this session is to provide a research-based framework and set of studies that illustrate ways to examine hybrid scoring models in live settings and provide data that can be referenced by others in the field.

Session Organizer:
Susan Lottridge, Cambium Assessment, Inc

Participants:
Implications of the Dependence of Hybrid Scoring Systems on Human Scoring Traditions
Michelle Boyer, Data Recognition Corporation

Examining Hybrid Scoring Results in One State Assessment Program
Susan Lottridge, Cambium Assessment, Inc

Examining Hybrid Scoring Results in a Multi-State Design
Corey Palermo, Measurement Incorporated

Discussant:
David Williamson, College Board

090. Interpreting COVID-19 Test Scores: Mode Effects and Missing Data

Coordinated Paper Session
2:30 to 4:00 pm PT
Westin San Diego Gaslamp: California Ballroom C

The COVID-19 pandemic has impacted every aspect of K-12 education in the United States. Available data from interim and summative assessments administered in the 2020-21 school year indicate that students' scores are lower this year relative to a "typical" school year, particularly in math. However, the pandemic has created several fundamental challenges to test score comparability that must be considered when using scores to understand that impact. In this panel, we discuss on-going research on the benefits and pitfalls of using assessment data collected during the 2020-21 school year to understand the impacts of the pandemic. Specifically, we focus on questions of remote/in-person test comparability and patterns of missing data using assessment data from two assessments (a) the NWEA MAP Growth math and reading tests in grades 3-8 and (b) a mandatory assessment of students' early literacy skills in Virginia. Through four papers, this coordinated session explores test score patterns during the pandemic, investigates how changes in test-taking populations may impact school-level inferences, and examines the impact of assessment mode. The implications of these issues for promoting fair and accurate interpretations of pandemic-era test score data are discussed.

Session Organizer:
Jonathan Schweig, RAND Corporation

Participants:
Learning During COVID-19: Reading And Math Achievement In The 2020-21 School Year
Megan Kuhfeld, NWEA

Understanding How Changes in Test-Taking Populations Impact School-level Inferences About COVID-19 Impacts
Jonathan Schweig, RAND Corporation

Were Tests Remotely Comparable In The 2020-21 School Year?
Patrick Meyer, NWEA; Megan Kuhfeld, NWEA

Early Literacy, Equity, and Test Score Comparability During the Pandemic
James Soland, University of Virginia

Discussant:
Andrew Ho, Harvard Graduate School of Education

091. Teaching Demonstrations

Demonstration Session
2:30 to 4:00 pm PT
Westin San Diego Gaslamp: Del Mar

Participants:
CTT's Blindness Regarding Item-Person Interactions in Anticipation of IRT
Ben Domingue, Stanford University
The CTT model offers no information about how a given person might respond to a given item. We use this fact to simulate data that meets the CTT model but where conventional estimates of reliability fail catastrophically. This didactic example is useful in motivating the innovations of IRT.

Creating and Implementing a Year-Long Seminar on Equity-Centered Assessment: Lessons Learned

Sara Finney, James Madison University; Brian C Leventhal, James Madison University

We share the process of developing a year-long, graduate-level, non-credit bearing seminar focused on anti-racist assessment and associated topics (e.g., QuantCrit, CRT). We share the mapping of student learning outcomes to the programming we created and deliverables expected from students. We share what was effective and what needs improvement.

Demonstrating a Lexical Parser for Improving the Readability of Multiple-Choice Questions

Magdalen Beiting-Parrish, CUNY Graduate Center; Jay Verkuilen, City University Of New York; Sydne McCluskey, CUNY Graduate Center; Howard Everson, CUNY Graduate Center

This demonstration will showcase the development and usability of a lexical parser for improving the readability of multiple-choice questions to better support English Language Learners in K-12 classrooms and higher education. This lexical parser is a web-based application that is easily accessed and demonstrated with any internet-enabled device.

Recruiting to the Measurement Profession by Employing an Undergraduate Internship

Brian C Leventhal, James Madison University; Kathryn Nicole Thompson, James Madison University

Faculty in measurement graduate programs consistently look for methods to recruit undergraduates to the profession. As a result, we have developed an undergraduate internship program with potential for wide-spread implementation due to availability of remote-work infrastructure. In this session, we will discuss strategies for implementation based on lessons learned.

Reliability challenge! A Game to teach CTT reliability in educational measurement

Sergio Araneda, University of Massachusetts Amherst

I will present a game in which players are asked to re-order two symmetrical deck of slides that present the results of two tests: one with low reliability and one with high reliability. Players need to use their advanced knowledges about reliability to solve this fun puzzle.

092. Practical Considerations in Transitioning IRT Calibration Software

Coordinated Paper Session
2:30 to 4:00 pm PT
Westin San Diego Gaslamp: La Jolla

When an organization decides to change IRT calibration software due to availability, cost, or feature limitations, psychometricians may be faced with a multitude of practical considerations. As psychometricians, our chief goal is test score validity; therefore, it is pertinent that software replacement be conducted in a data-driven way to ensure a smooth, accurate, and defensible transition. The papers in this coordinated session discuss three unique testing programs in the midst of replacing their IRT calibration software. Each testing program is unique in the IRT model applied, target examinee population, and software that is being retired and newly implemented. Although the investigations are unique, the findings and practical implications are common to operational psychometricians across the field.

Session Organizer:

Matthew Swain, American Board of Internal Medicine

Chair:

Aaron Myers, University of Arkansas

Participants:

Replacing PARSCALE with SAS PROC IRT

Matthew Swain, American Board of Internal Medicine; Derek Sauder

Comparing BILOG-MG and SAS IRT Calibration Procedures

Ying Jin, Association of American Medical Colleges; Hye-Jeong Choi, Human Resources Research Organization; Marc Kroopnick, Association of American Medical Colleges; Bethany Bynum, HumRRO

On Sensitivity of SAS PROC IRT(MCMC) and R package ltm in 3-PL Models

Hye-Jeong Choi, Human Resources Research Organization; Bethany Bynum, HumRRO

Comparison of Calibration Methods for CAT Pool Development

Emre Gonulates, Human Resources Research Organ; Matt Trippe, Human Resources Research Organization; Ted Diaz, Human Resources Research Organization; Olga Golovkina, HumRRO; Mary Pommerich, Defense Personnel Assessment C; Daniel Segall, DMDC

Discussant:

Erin Banjanovic, Pearson

093. Applications of Bias, Fairness, DIF

Paper Session
2:30 to 4:00 pm PT
Westin San Diego Gaslamp: Santa Fe

Participants:

Investigating Item Parameter Drift of Sense of Belonging Measure
Tai Do, University of Minnesota; Rik Lamm; Michael C. Rodriguez, University of Minnesota
Item parameter drift (IPD) and differential item functioning (DIF) was investigated for a measure of sense of belonging administered to adolescents. Results indicated negligible DIF, however, differences in endorsement of the measure and one item were extant for students of color compared to White students.

Item Format Gender Bias: New Evidence and Considerations for State Accountability Testing
Benjamin R. Shear, University of Colorado Boulder
This paper compares gender DIF across multiple-choice (MC) and constructed response (CR) items for three recent PISA administrations. The analyses investigate associations between item features and DIF and discuss implications for state accountability testing, where the proportion of MC and CR items varies widely across states and grades.

Time to Add Ethical Standards for Using Process Data in Educational Assessment?
Fazilat Siddiq, University of South-Eastern Norway; Damian Murchan, Trinity College Dublin
Computerized assessments can generate process data that record examinees' digital traces. Whereas advances in analyzing such data accelerate, ethical and legal concerns persist. This study relates the paucity of information about ethics in process data studies to existing test standards, proposing how standards might be revised to address the concerns.

Discussant:
Mark Johnson, Cognia, Inc.

094. Advanced Methods Using Large Scale Assessment Data

Paper Session
2:30 to 4:00 pm PT
Westin San Diego Gaslamp: Plaza

Participants:

Examining the Effects of Speededness on Group Ability Estimation in NAEP
Young Yee Kim, American Institutes for Research;Xiaying Zheng, American Institutes for Research; Xiaoying Feng, American Institutes for Research; Yi Yang, Teachers College Columbia University
Research has shown test speededness can bias IRT model parameter estimates and thus affect test validity. This study intends to evaluate the effects of test speededness on IRT estimation and group ability estimation with a simulation study and further provide guidelines to measure test speededness using simple missing rate measure.

Generating Synthetic Large-Scale Assessment Data
Sinan Yavuz, American Institutes for Research; Ting Zhang, American Institutes for Research; Paul Bailey, American Institutes for Research
Large-scale assessment datasets are frequently used for educational research, but often restricted access to data makes it difficult to do research. Synthetic data can help expand research with large-scale assessments by allowing the exploration of new methodologies. We generated a NAEP-like dataset with two different methods and compared the results.

A Network Psychometrics Approach for Detecting Item Wording Effects Across Subgroups
Hatice Cigdem Bulut, Cukurova University; Okan Bulut, University of Alberta
The goal of this study was to propose a network modeling approach for detecting wording effects. Using two measures in TIMSS 2019 administered to 75,972 students from 8 countries, we employed Explanatory Graph Analysis and network model trees to examine the presence of wording effects across countries and age groups.

Proper Use of Test Scores from A Large Scale Assessment
Salih Binici, Florida Department of Education; Yachen Luo
This study examines consequences of model misfit and measurement error on reporting outcomes for a large scale assessment. It investigates whether ignoring model misspecification and measurement error has any practical impact on reported scale scores for parents and teachers, also their secondary use in statistical analyses to inform policy makers.

Discussant:
Qiwei He, Educational Testing Service

096. De-Centering Whiteness in Assessment Practices and Products

Organized Discussion
4:15 to 5:45 pm PT
Westin San Diego Gaslamp: California Ballroom A

Standard 3.2 in the Standards for Educational and Psychological Assessment asserts that "Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests' being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics" (p.64). This standard is justified on the basis that "unnecessary linguistic, communicative, cognitive, cultural, physical, and/or other characteristics in [tests] can impede some individuals in demonstrating their standing on intended constructs" (p.63). In practice, this removal of "unnecessary" cultural and linguistic characteristics often results in instruments that reflect the dominant culture of whiteness (often unconsciously) while excluding or erasing other groups. What is the cost of that erasure? What are the potential benefits of undoing it? How can the assessment community learn about and respond to this reality? Join us for a discussion of the importance of de-centering whiteness in testing, and practical solutions to building anti-racist, culturally relevant, and culturally sustaining testing products. Our panelists will think concretely about ideologies and methodologies that are underutilized in our field yet have the potential to bring educational justice to historically slighted groups. The discussion will be framed as an interrogation that supports action, not optics.

Session Organizer:
Molly Faulkner-Bond, WestED

Moderator:
Darius D. Taylor

Presenters:
Kristen Huff, Curriculum Associates
Michael C. Rodriguez, University of Minnesota
Paula Groves Price, North Carolina Agricultural and Technical State University

097. At the Crossroads of Psychometrics and Causal Inference

Coordinated Paper Session
4:15 to 5:45 pm PT
Westin San Diego Gaslamp: California Ballroom B

Causal research in education has made a lot of progress since the 1990s, with big improvements in quasi-experiments and Randomized Control Trials (RCTs) being re-established as a mainstream method (e.g., Hanley et al., 2016; Kathryn & Paul,2019; Wiseman, 2010). Substantial attention has been paid to issues of study design and analysis, but almost no attention has been given to the construction and design of measures of learning outcomes and its psychometric properties, particularly, when re-searchers design measures themselves rather than use state or district exams as outcomes (Olsen et al., 2013; Bloom et al., 2014; Shadish et al., 1998). With this symposium we aim to shed light on the state of knowledge regarding interplay between the psychometric properties of measures and the associated causal estimates (within the US and international contexts). The first paper documents the poor quality of measures of learning outcomes in RCTs in over 150 trials using both CTT and IRT methods. The second paper utilizes DIF analysis to evaluate whether observed treatment effects are due to generalized gains or gains that largely manifest with respect to specific items. The final paper uses CTT and a causal framework to quantify the effect of retesting on test standards.

Session Organizer:
Masha Bertling

Chair:
Masha Bertling

Participants:
Do We Know What and How to Test? Improving Measures of Student Achievement in Development Economics
Masha Bertling

Differential Item Functioning Across Randomized Controlled Trials in Educational Settings
Ishita Ahmed, Stanford; Ben Domingue, Stanford University

Effect of Retesting Policy on Exam Standards and Standard Setting
Sophie Litschwartz

Discussant:
Ben Domingue, Stanford University

098. Applications of Response Time in Measurement

Coordinated Paper Session
4:15 to 5:45 pm PT
Westin San Diego Gaslamp: California Ballroom C

The measurement community has been witnessing a persistent shift to computerized and online testing for decades. The COVID-19 pandemic has accelerated this shift as many testing programs scrambled to switch to computerized and/or online testing. Despite the advantages computerized/online testing introduces, many challenges have surfaced, such as test security (due to less controllable testing environment), interruptions to test taking (due to unstable access for instance), test fairness (possibly introduced by unequal access or lack of accommodations), and lack of feedback (due to fewer opportunities to interact with teachers), to name a few. Meanwhile, many researchers had to resort to online data collection as well. This raises the question of data quality too. In this session, we attempt to address these issues in high-stakes and low-stakes testing environments, by using response times that are collected during computerized/online testing process.

Session Organizer:
Ying Cheng, University of Notre Dame

Participants:

Joint Bi-factor Modeling of Responses, Response-Time, and Answer Changes for Cognitive Diagnosis
Hong Jiao, University of Maryland; Todd Zhou, Univeristy of Maryland; Yishan Ding, University of Maryland

Using response time for compromised item detection
Cheng Liu; Kyung (Chris) Han, Graduate Management Admission Council; Jun Li, University of Notre Dame

Person-Explanatory Response Time Model for a Low-Stakes Assessment
Daniella Rebouças-Ju, University of Notre Dame; Ying Cheng, University of Notre Dame

A sequential Bayesian changepoint detection procedure for aberrant behaviors
Jing Lu, Northeast Normal University; Chun Wang, University of Washington; Jiwei Zhang, School of Mathematics and Statistics, Yunnan University

Detecting Differential Item Functioning Using Response Times
Qizhou Duan, University of Notre Dame; Ying Cheng, University of Notre Dame

Discussant:
Sandip Sinharay, Educational Testing Service

099. Cheating Detection: A Collaborative Case Study using IT Certification Exams

Coordinated Paper Session
4:15 to 5:45 pm PT
Westin San Diego Gaslamp: Del Mar

Cheating damages the integrity of a testing program and can cause testing organizations significant losses. Security breaches can arise from individuals memorizing and sharing items, the concerted efforts of a test preparation company to harvest items and teach them to their customers, and answer copying or collusion among examinees during a testing event. Without proper detection, these types of cheating could remain undetected until their presence becomes significant enough to threaten test-score validity. It is crucial for a test sponsor to accurately identify cheaters and invalidate their scores to effectively deter cheating behaviors. However, many cheating detection techniques developed so far are based on complicated mathematical models and extensive ad-hoc data analyses and thus cannot be practically conducted on daily basis. Therefore, we propose a collaborative exercise in which five independent research groups each propose a method that could help effectively and efficiently detect cheaters in the operational setting. Each group will use the same data from two linear fixed-form IT certification exams with known security breaches. The five approaches will be comparatively evaluated regarding their accuracy in detecting cheating and feasibility to implement.

Session Organizer:
Anjali Weber, Amazon Web Services (AWS)

Participants:

Varying item parameters and collusion detection
Kirk Becker, Pearson; Paul Edward Jones, Pearson VUE

Simultaneous Estimation of Compromised Items and Examinees with Item Preknowledge
Cengiz Zopluoglu, University of Oregon

A practical application of response similarity
Russell Smith, Alpine Testing

Identifying Anomalous Response Patterns with Multinomial Logistic Regression
Jennifer Davis, Amazon Web Services (AWS)

Machine Learning Based Profiling in Test Fraud Detection
Huijuan Meng, Amazon Web Services (AWS)

Discussant:
James Wollack, University of Wisconsin

100. Research Blitz

Research Blitz Session
4:15 to 5:45 pm PT
Westin San Diego Gaslamp: La Jolla

Chair:
Leah Feuerstahler, Fordham University

Participants:

Assessment Use for Accountability and Informing Instruction: Impacts on Teachers and Students
Carina M. McCormick, Buros Center for Testing
The study examines the use of standardized assessments in schools and its relationship to views of teachers and students in the school, using data from U.S. public schools sampled in the 2018 PISA administration. Relationships with teacher satisfaction and student-reported SEL characteristics and teacher support are evaluated using multilevel models.

Examining Differential Rates of Progress Along an Early Childhood Learning Progression
Joshua Sussman, UC Berkeley; Mark Wilson, UC Berkeley
We describe findings from a new psychometric method related to differential item functioning designed for use with longitudinal data collected from the perspective of a learning progression. We modeled the emergence of learning differences between groups of children and identified specific learning subdomains that contributed most to the observed differences.

Exploring Learning Pathways in a Critical Online Reasoning Assessment among graduates
Susanne Schmidt; Olga Zlatkin-Troitschanskaia, Johannes Gutenberg University; Kevin Shenavai, Johannes Gutenberg University Mainz; Bültmann Ann-Kathrin, Johannes Gutenberg University Mainz
Within the digital Critical Online Reasoning Assessment graduates are asked to follow one URL (e.g. a twitter link) and judge whether the information is reliable by conducting a free web search. Written statements from the participants, along with event data, enables us to visualize and evaluate their learning pathways.

Issues in Instructional Sensitivity in an Era of Remote Learning
Audra Kosh, Edmentum; Jinah Choi, Edmentum, INC.
This conceptual paper presents issues related to instructional sensitivity in an era of remote learning, such as how to define instruction when personalization and opportunity to learn varies due to heavily-utilized educational technology tools during school closures from the COVID-19 pandemic. Empirical examples from an adaptive test are included.

Measuring group collaboration using process data in an online computer-based assessment environment
Nafisa Awwal, University of Melbourne; Mark Wilson, UC Berkeley; Zhonghua Zhang, University of Melbourne
The authors adapt a multidimensional framework for collaborative problem-solving to assess collaboration in groups and individuals interacting in computer-games. Process data from group activities is coded into behavioural indicators to indicate collaboration. The results of the empirical data provided initial validity evidence of the framework in measuring group collaboration during solving problems.

Moving Toward Online Assessment: Issues, Equity and Exemplars
Stanley N Rabinowitz, Pearson
Many assessment programs have made the move to online. This paper examines: • What are the challenges facing exemplar programs moving to an online assessment? • How did these exemplar programs—selected from the U.S.A. and international—address these challenges? • What is the impact on equity and inclusion as programs move online?

The Relationship between Skip Year and Traditional Student Growth Percentiles
Jessalyn Smith, DRC; Scott Li, DRC
SGPs tend to be used for populations and content areas that are stable across time, with yearly administrations. The purpose of this study is to use longitudinal data from a large-scale assessment to examine consistency of student growth percentiles and targets when cohorts of students skip an administration.

101. A Content-Referenced Approach to the Interpretation of Growth

Coordinated Paper Session
4:15 to 5:45 pm PT
Westin San Diego Gaslamp: Santa Fe

We present a novel approach to ascribing content-referenced meaning to students' scores using Curriculum Associates' i-Ready Diagnostic assessment software tool for illustration. This approach, called content-referenced growth, has four ingredients: (1) a learning progression; (2) a vertical scale; (3) item mapping; and (4) graphical visualizations. The goal of content-referenced growth is to support interpretations of students' scores relative to both the status of their understanding at one point in time, and their growth in understanding across points in time, relative to the content contained in the assessment. This coordinated session first describes the motivation for content-referenced growth as part of i-Ready's overall approach to quantifying and reporting growth. We then describe how each of the above "ingredients" forms an integral part of the content-referenced growth concept. The development and analysis of a mathematics learning progression in fractions is the basis for an extended example of what content-referenced growth would look like in practice. Analysis of the i-Ready vertical scale investigates the extent to which i-Ready's current scoring practices can support the types of interpretations required for content-referenced growth. Finally, a visualization shows how the other ingredients can come together to support meaningful diagnostic inferences about student status and growth.

Session Organizer:
***Derek Briggs*,** University of Colorado

Participants:
An Overview of the Role of Growth in the i-Ready Diagnostic Interim Assessment
Laurie Davis, Curriculum Associates; Daniel Mix, Curriculum Associates

The Development of a Learning Progression for Fractions with a Theoretical and Empirical Basis
Sarah Wellberg, University of Colorado, Boulder

Calibrating and Validating the Uses of a Vertical Scale in a Computerized Adaptive Setting
Sanford Student, University of Colorado Boulder

Visualizing Content-Referenced Growth Using Reference Items
Derek Briggs, University of Colorado

Discussants:
Richard Melvin Luecht
Frederick Peck, University of Montana

102. Connecting Ambitious Teaching and the Formative Assessment Process

Coordinated Paper Session
4:15 to 5:45 pm PT
Westin San Diego Gaslamp: Plaza

Unlike benchmark and summative assessments that provide measurements of learning (after intervals of instruction or the conclusion of instruction), assessment for learning through formative assessment should be an ongoing process enacted through classroom actions and interactions among peers and between teachers and students. More specifically, formative assessment is a process that involves careful planning, goal setting, targeted instruction, eliciting evidence of student learning, providing feedback from a variety of sources (teachers, students, self), and responding to this feedback through both teachers' instructional actions and students' learning decisions. For this coordinated paper session, we tap into data from a multi-year research project on the enactment of the formative assessment process in Michigan. Contributors to the papers have had a deep and sustained involvement in the Formative Assessment for Michigan Educators (FAME) program. Furthermore, the contributors have a variety of backgrounds—teachers, school/district administrators, state administrators, and university researchers—that will provide a thorough understanding of the formative assessment process and how it connects to ambitious teaching.

Session Organizer:
John Lane

Chair:
***Ellen Vorenkamp*,** Michigan Assessment Consortium

Participants:
What is the Formative Assessment Process?
Kristy Walters, Corunna Public Schools

The Link between the Formative Assessment Process and Recent calls for More Ambitious Teaching
Margaret Heritage

How Effective Use of Formative Assessment Practices Can Deepen Disciplinary Understandings (and Vice Versa)
Tara Kintz, Michigan Assessment Consortium; Amelia Gotwals, Michigan State University

Understanding how Teachers Elicit and Respond to Student Understanding through Questions
John Lane

Teacher innovation in response to current educational needs: Factors that support teacher learning about the formative assessment processes
Tara Kintz, Michigan Assessment Consortium; Ellen Vorenkamp, Michigan Assessment Consortium

Discussant:
Edward Dean Roeber, Michigan Assessment Consortium

103. NCME Yoga

6:30 to 7:30 am PT
Westin San Diego Gaslamp: Garden Terrace

104. Sociocultural Context of Assessment

Organized Discussion
8:00 to 9:30 am PT
Westin San Diego Gaslamp: California Ballroom A

This session will focus on the multiple facets that define sociocultural context in assessment. The importance of a sociocultural perspective in assessment has been recognized in the recently approved NAEP Reading Framework and highlighted in state assessments. Conventional approaches to addressing diversity of student populations and fairness in assessment pay attention to the content, wording, or format of items. While necessary, such approaches are not guided by principled practices that allow detailed, systematic consideration of the ways in which culture and social context shape the mind and influence the ways in which students interpret items and respond to them. Improved practices should promote equity and fairness in assessment. Issues relevant to this discussion include: What are the multiple facets of sociocultural context in assessment? How is sociocultural context relevant to assessment development, measurement modeling and score interpretation? How should assessment practices be modified to ensure proper consideration of sociocultural context and address inequities in learning outcomes? How can research, practice and policy be more effectively grounded in learners' social and cultural experiences? This interactive organized discussion will bring together researchers, practitioners and policy makers with extensive experiences in assessment and facilitate discussions among the panelists and with the audience.

Session Organizer:
Kadriye Ercikan, Educational Testing Service

Presenters:
Peggy Carr, Department of Education
Maria Araceli Ruiz-Primo
Linda Darling-Hammong, Learning Policy Institute
Jim Gee, Arizona State University
P David Pearson, University of California - Berkeley

Discussant:
Guillermo Solano-Flores, Stanford University

105. Components of a Well Balanced Assessment System

Organized Discussion
8:00 to 9:30 am PT
Westin San Diego Gaslamp: California Ballroom B

This session with begin by setting the context for the session and the four presentations that follow. School districts have a wide range of stakeholders with varying information needs which can be supported by different kinds of assessments. The design of an assessment intended to support school accountability determinations will be quite different from a classroom assessment intended to inform how the remainder of a lesson will play out. Since stakeholders require different types of information to support educational decision-making students will participate in multiple types of assessment events within a given year. However, it is important that assessments are not used for purposes to which they are not well suited. Across the presentations in this session, the presenters will focus on how assessment directors can support classroom assessment that primarily is intended to meet the needs of teachers and students.

Session Organizer:
Darin Kelberlau, Millard Public Schools

Moderator:
Charlotte Gilbar, Natrona County School District

Presenters:
Caroline Wylie, ETS
Alison Bailey, UCLA
Heidi Andrade, University at Albany
Erika Landl, Center for Assessment
Kendra Pullen, Caddo Public Schools

106. eBoard Session

Electronic Board Session
8:00 to 9:30 am PT
Westin San Diego Gaslamp: San Diego Ballroom

Participants:

What Does GRM Miss? Revisiting ICCS Civic Participation Items Using Network Analysis

Nicolas Riveros Medelius, Harvard Graduate School of Education; Eva Flavia Martinez Orbegozo, Harvard Graduate School of Education

Using data from the 2016 wave of the ICCS we construct networks of student co-reported participation in civic activities and generate student participation scores (SNA-scores). We compare these scores with scores estimated from a graded response model (GRM-scores) and find SNA-scores capture a highly correlated but distinct object of measurement.

Posterior Predictive Model Checking of the Hierarchical Rater Model

Nnamdi Ezike, University of Arkansas; Allison Ames Boykin, University of Arkansas

We risk making invalid inferences if a posited model does not fit the data. This study investigates the performance of the Bayesian posterior predictive model checking approach in detecting model-data misfit of the hierarchical rater model. Using different discrepancy measures, we evaluated misfit at the test, item, and rater levels.

Comparison of MIMIC Model and Traditional DIF Detection Methods under Missing Data

Onur Demirkaya, University of Illinois at Urbana-Champaign; Jinming Zhang, University of Illinois at Urbana-Champaign

This simulation study investigated how three methods for dealing with missing data (listwise deletion [LW], two-way imputations [TW], multiple imputations [MI]) impacted the performance of DIF detection in the MIMIC model, logistic regression, and SIBTEST under three mechanisms of missingness.

Assessing Middle School Student Computational Thinking in an Immersive Game Environment

Paul Foster, Southern Methodist University; Ching-Yu Tseng, Southern Methodist University; Lawrence J. Klinkert, Southern Methodist University; Elizabeth L. Adams, Southern Methodist University; Leanne Ketterlin Geller, Southern Methodist University; Eric C. Larson, Southern Methodist University; Corey Clark, Southern Methodist University

Using a commercial game such as Minecraft as a platform may promote learning by increasing student engagement. The challenge is meeting content standards. This poster illustrates how we identified the evidence needed to make inferences within an immersive game environment about a student's knowledge, skills, and abilities.

Modeling a summative assessment for a multidimensional learning progression

Perman Gochyyev, University of California, Berkeley; Jerred Jolin, Eastern Oregon University; Mark Wilson, UC Berkeley

Learning progressions (LPs) are often multidimensional due to the scope of the learning represented. We investigated qualities of different approaches to estimate student ability at the overall macro level, in addition to each underlying dimension (or construct), in the context of a college-readiness assessment in problem solving using mathematics (PSM).

Pros and Cons of Technology-enhanced in a Digital Literacy Assessment

Qianqian Pan, The University of Hong Kong; Yuxiao Zhang, Purdue University; Patrick Lam, The University of Hong Kong

Technology-enhanced items (TEIs) have particular strengths in measuring digital literacy (DL), including increasing test-takers' engagement levels. However, TEIs also pose threats to validity, including introducing nuisance factors. This study evaluated TEIs in a territory-wide DL assessment. Results confirmed that TEIs show fewer rapid-guessing behaviors but introduced a measurable dimension.

Measurement Invariance of English Proficiency Test on U.S. and International ELLs

Sakine Gocer Sahin, WIDA at UW-Madison; Kyoungwon Lee Bishop, WIDA at University of Wisconsin Madison; Sinan Yavuz, American Institutes for Research

In this study, we will examine measurement invariance (i.e., configural, metric, scalar, and strict invariance) of an English Proficiency Test and each domain on U.S. and International ELLs. Two samples composed of 3,682 ELLs in the U.S. and 2,177 ELLs in other countries will be analyzed.

Effectiveness of Speededness Detection Methods under Different Speededness Scenarios

Shichao Wang, ACT; Ann Arthur, ACT; Dongmei Li, ACT

This study offers an intensive investigation of the effectiveness of three speededness detection methods, including change-point analysis, mixture-IRT model with ordinal constraints, and a response time threshold method, using both real-world experimental data and simulated data generated from different speededness models.

Comparing Methods for Detecting Prior-Data Disagreement in Bayesian Structural Equation Modeling

Sonja D Winter, University of Missouri; Sarah Depaoli, University of California, Merced

The choice of prior specification plays a vital role in any Bayesian analysis. Prior-data disagreement occurs when the researcher's prior knowledge is not in agreement with the evidence provided by the data. This study compares three methods for detecting prior-data disagreement through a simulation design.

Relationships between students' college major interests and ACT score

Sunhyoung Lee, University of Nebraska-Lincoln; James Bovaird, University of Nebraska-Lincoln; Hongwook Suh, Nebraska Department of Education

This study investigates how students' college major interests and their assurance of their choice of the major are related to the performance on the ACT score. Three-level multilevel modeling is employed for the hierarchical data. The results and discussion of the analyses will provide practical implications regarding the ACT score.

Student Perceptions of the No-Choice Shift to Online Learning

Tabasom Fayaz, Kwantlen Polytechnic University; Shayna Rusticus, Kwantlen Polytechnic University; Dianne Crisp, Kwantlen Polytechnic University; Wayne Podrouzek, Kwantlen Polytechnic University

In an online survey of perceptions of learning online after 10 month's immersion, university students (n=185) indicated that faculty flexibility and responsiveness supported their learning, but that challenges in terms of lack of direct interaction and excessive screen time had a negative impact on their emotional responses to this environment.

Opportunity-to-Learn as Explanatory Sources of Differential Gender Math Performance: A Multilevel Framework

Thao Vo, Washington State University; Brian French, Washington State University; Shenghai Dai, Washington State University

Opportunity-to-learn (OTL) indicators are explored as sources of gender-related differential item functioning (DIF) using multilevel random-intercept and random-coefficient models. Results indicate that the 2019 Trends in International Mathematics and Science Study eighth-grade mathematics assessment displayed DIF and school-level OTL indicators accounted for proportions of the variance related to DIF.

Scale Transformation Methods to Link the Multistage Test

TsungHan Ho, ETS

Simultaneous linking and fix item parameter calibration are two methods that appeared to be more robust to fluctuations in volumes for linking items in a multistage test. The performance of both methods is evaluated by the comparison with Stocking-Lord procedure in terms of the item parameter recovery across test conditions.

A Bayesian Approach to Scaling Panelist-Item Interactions during Standard Setting

William Skorupski, Data Recognition CORP; Joseph Fitzpatrick, NWEA

A Bayesian method for evaluating cutscores from Angoff standard setting is presented. Panelists' judgments about the probability of item success for minimally competent candidates are used to develop a modified IRT model of expected examinee behavior. The approach is presented using real and simulated data.

The Research-based Early Mathematics Assessment: Validation and Score Equating Studies

Yixiao Dong, University of Denver; Douglas Clements, University of Denver; Crystal Day-Hess, University of Denver; Julie Sarama, University of Denver; Denis Dumas, University of Denver

This research reports on three empirical studies (validation, cross-observer validity, and score equation) that encompass the development and validation of the Research-based Early Mathematics Assessment-Short Form (REMA-SF), an instrument that measures the early mathematical competency of children from 3 to 8 years of age.

Making construct-irrelevant variance relevant: the case of resilience and other possibilities

Yoav Bergner; Susu Zhang, University of Illinois at Urbana-Champaign

It is sometimes possible to explicitly model the "construct-irrelevant" sources of variance in testing. These models are multidimensional and can be finicky. We describe one fruitful approach to modeling resilience alongside ability in tests with multiple chances to answer. We also suggest other directions for inclusive approaches to test scoring.

Automated Essay Scoring Using Neural Network Model with Hybrid Features

YoungKoung Kim, College Board; Tim Moses, College Board

The present study proposed a new machine learning technique that uses hybrid features, i.e. combinations of handcrafted features and deep learning sentence features extracted from the BERT model for automated essay scoring. The results showed that the proposed model greatly improved the automated essay score agreements with human rater scores.

107. Innovative Assessment Systems: Informing the Future of Educational Measurement

Coordinated Paper Session
8:00 to 9:30 am PT
Westin San Diego Gaslamp: Del Mar

As the conference theme suggests, the field of K-12 educational measurement is rapidly changing. Impacts of the COVID-19 pandemic, states pursuing innovative assessment systems under the U.S. Department of Education Innovative Assessment Demonstration Authority, and advances in technology are changing the way we approach K-12 educational assessment. New assessment models are increasingly being explored to provide teachers with actionable information about student learning, and states are considering alternate approaches to meeting accountability requirements. Innovative scoring models allow programs to provide timely results throughout the year. In light of these advancements, we describe four innovative assessment programs in varying stages of design and implementation. We discuss the innovations they offer and share lessons learned from each program that can help inform the future of innovative assessment systems and K-12 educational measurement.

Session Organizer:

Amy Clark, ATLAS: University of Kansas

Chair:

Amy Clark, ATLAS: University of Kansas

Participants:

Lessons Learned from Stackable, Instructionally-Embedded, Portable Science Assessments

James Pellegrino, University of Illinois at Chicago

Lessons Learned from Navvy
Laine Bradshaw, University of Georgia

Lessons Learned from Louisiana Educational Assessment Program
Chanda Johnson, Louisiana Department of Education

Lessons Learned from Dynamic Learning Maps Alternate Assessment System
Meagan Karvonen, University of Kansas; Amy Clark, ATLAS: University of Kansas

Discussant:
Scott Marion, National Center for the Improvement of Educational Assessment

108. PISA Applications

Paper Session
8:00 to 9:30 am PT
Westin San Diego Gaslamp: La Jolla

Participants:

Alternatives to Weighted Item Fit Statistics for Establishing Measurement Invariance
Montserrat B Valdivia Medinaceli, Indiana University Bloomington; Sean Joo, University of Kansas; Leslie Rutkowski, Indiana University; Dubravka Svetina Valdivia, Indiana University
Root mean square deviation (RMSD), currently used to establish measurement invariance in PISA, is not well suited for detecting DIF in low-performing countries. We propose new alternative approaches to compute RMSD: equal weight, item information weight, and b-norm weight. The performances of all methods were evaluated via a simulation study.

An Investigation of Item, Examinee, and Country Correlates of Rapid Guessing on PISA
Joseph A. Rios, University of MinnesotA; James Soland, University of Virginia
The objective of the present study was to investigate item-, examinee-, and country-level correlates of rapid guessing (RG) in the context of the 2018 PISA science assessment. Data were analyzed from 267,148 examinees across 71 countries using multiple two-level cross-classified hierarchical linear models.

Dimensionality Structure of Large Scale Assessments. Case of PISA
Artur Pokropek, IFIS PAN; Francesca Borgonovi, University College London; Organisation for Economic Co-operation and Development; Piotr Koc, Polish Academy of Sciences; Gary Neil Marks, The University of Melbourne
Psychometricians working on LSAs typically specify latent ability domains as distinct and correlated constructs. We examine data from the 2018 PISA assessment in 37 countries using IRT bifactor-modeling. We identify strong evidence of unidirectionality of the PISA assessment which, together with additional analyses, suggest that PISA achievement scores mostly reflect general cognitive ability rather than domain specific abilities.

Investigating Students' Response Times in International Survey Assessments
Mina Lee, University of Massachusetts, Amherst; Frederic Robin, ETS; Hyo Jeong Shin, Educational Testing Service; Hongwen Guo, Educational Testing Service
We investigated how to best fit the Bayesian hierarchical model for speed and accuracy (van der Linden, 2007; Entink et al., 2009) to PISA 2015 Science. The model's assumptions as well as the effects of item types on response times were examined. Aberrant RT patterns were further investigated.

Identifying problem-solving solution patterns using network analysis of process data
Maoxin Zhang; Bj□rn Andersson, Centre For Educational Measure
Process data from educational assessments provide important information about how students answer cognitive items. We propose to visualize the process data as a network, define network-derived variables to extract essential information, and identify student solution patterns through cluster analysis. The method is applied to a problem-solving task from PISA 2012.

Discussant:
Michael Brannen Bunch, Measurement Incorporated

109. COVID-19 Impact

Paper Session
8:00 to 9:30 am PT
Westin San Diego Gaslamp: Santa Fe

Participants:

COVID-19 impact on Group Invariance Property of Equating
Dong-In Kim, Data Recognition Corporation; Marc W Julian, DRC; Joanna Tomkowicz, Data Recognition Corporation; Pamela Hermann, Data Recognition Corporation
Student learning has been unevenly impacted by circumstances related to the COVID-19 pandemic, which in turn could have negatively affected one of the critical equating properties, the group invariance. This study examines the group invariance property when pre-equating is applied to large-scale assessments in spring 2021 administration.

Effects of COVID-19 on Student Achievement in Large Scale Assessments

Aurore Yang Phenow, Data Recognition Corporation; Dong-In Kim, Data Recognition Corporation; Keith Boughton, Data Recognition Corporation

This study utilized multilevel mixed effects modeling to assess the effect of student and school level predictors on achievement in large scale assessments. Coefficients from 2021 models were compared to 2019 models to understand and identify how changes in instruction due to COVID-19 may have impacted student learning.

How Much They Have Lost and How Long to Recover?

Chalie Patarapichayatham, Southern Methodist University; Victoria Locke, Istation

This study investigates how much K-5 students across the US have lost their reading and math achievement levels during the COVID years and how long it would take for them to recover from the COVID-19 learning loss.

Discussant:

Karla Egan, EdMetric, LLC

110. Focus on Equity and Social Justice

Paper Session
8:00 to 9:30 am PT
Westin San Diego Gaslamp: Plaza

Participants:

Argument-based Fairness in Educational Measurement, with an AI-Enhanced Assessment Application

Anne Corinne Huggins-Manley, University of Florida; Brandon M. Booth, University of Colorado Boulder; Sidney D'Mello, University of Colorado Boulder

The purpose this study is two-fold: (1) introduce an argument-based fairness approach that contrasts with current practices of arguing fairness in the form of rebuttals to validity arguments, and (2) apply an illustrative case-study of the argument-based fairness approach in an AI-enhanced assessment.

Black Lives Matter's Influence on the Black White Achievement Gap

Charles Secolsky; Randall Richards III, New Jersey Association of Student Financial Aid Administrators

African-American students are disadvantaged in educational achievement as evidenced by black-white Achievement Gap. Black Lives Matter motivates blacks. This study examines whether Achievement Gap narrowed. A survey based on interviews sheds light on whether BLM influences this narrowing. Trend analysis ANCOVA was performed SAT or Accuplacer covariates from 2016-2021.

Measuring Equitable Mathematics Instruction: An Evaluation of Scores Using Generalizability Theory

Elizabeth L. Adams, Southern Methodist University; Anne G. Wilhelm, Southern Methodist University; Rachael N Becker, Southern Methodist University; Jonee Wilson, NC State University; Temple A. Walkowiak, NC State University; Anna Thorp, NC State University; Tiffini Pruitt-Britton, Southern Methodist University; Danielle Moloney, NC State University; Natalia Yanez-Castillo, NC State University

We evaluated the score stability of the Equity Rubrics, an observational measure of equitable instructional practices in elementary and middle grades mathematics classrooms. Five raters independently scored 60 video-recorded lessons for 24 teachers. Using generalizability theory, we decompose score stability into potential sources of variation.

Turning the Page: Leveraging Opportunity to Learn Data through Community Collaboration

Kerry Englert, Seneca Consulting, LLC; Pohai Kukea Shultz, University of Hawaii at Manoa; Jessica Allen, Seneca Consulting

The Kaiapuni Assessment of Educational Outcomes team used the challenges presented by the pandemic to collaborate with the Hawaiian immersion school community to develop relevant opportunity to learn (OTL) surveys. This paper will describe how surveys were built and data reported that are valuable to and supportive of communities.

Discussant:

Susan Lyons, Lyons Assessment Consulting

111. Bifactor Models

Paper Session
8:00 to 9:30 am PT
Westin San Diego Gaslamp: California Ballroom C

Participants:

Integrating Bifactor Models into G-Theory Frameworks

Walter Vispoel, University of Iowa; Hyeryung Lee, University of Iowa; Guanlan Xu; Hyeri Hong

We integrated bifactor models into multi-occasion G-theory SEM frameworks to allow for expanded score consistency indices; separation of systematic variance into general and group factor effects and measurement error into transient, specific-factor, and random-response effects; and generalization of results to broader domains from which items and occasions were sampled.

Factors That Affect the Empirical Identification of the Bifactor IRT Model

Wenya Chen, Loyola University Chicago; Ken Fujimoto, Loyola University Chicago

An empirical non-identification issue arises with the bifactor IRT model when an item's discriminations on the general and specific dimensions are similar. However, how similar the discriminations have to be and whether the similarity depends on sample size and strength of item discrimination is unknown, which our study addresses.

Integrating Cognitive and Non-Cognitive Data to Assess Student Engagement: Bi-Factor Joint Modeling

Hong Jiao, University Of Maryland; Todd Zhou, Univeristy of Maryland; Xin Qiao; Xiaoyu Chen, Future Education Foundation

This study proposes bi-factor joint modeling of item responses and response time in assessment, and survey item responses in questionnaires to assess student engagement by integrating cognitive and non-cognitive data in the 2018 PISA assessment and the questionnaires for the USA. Model parameter estimation and model performance are investigated empirically.

Measurement Alignment of Bifactor Model

Seohee Park, American Board of Internal Medicine

This study extends measurement alignment into a bifactor model. Although measurement alignment is used in various fields, applicable models are limited. Considering requests of application to the bifactor model, this study introduces procedures of measurement alignment for the bifactor model and demonstrates the procedures through empirical data and simulation analyses.

Discussant:

Jonathan Weeks, Educational Testing Service

112. Admissions Testing, Adverse Impact, and the Responsibility of the Testing Industry

Organized Discussion
9:45 to 11:15 am PT
Westin San Diego Gaslamp: California Ballroom A

Negative consequences associated with educational tests are always undesirable, and the responsibility for evaluating and remediating such consequences are contentious issues within the educational measurement community and the testing industry. The focal article on college admissions testing and corporate responsibility published in Educational Measurement: Issues and Practice, and the commentaries published in response to the article, illustrate different notions of what the testing industry can do to address the effects of systemic racism that result in large score gaps on admissions tests, and hence adverse impact in college admissions. This symposium fosters a dialogue on this issue by embracing different perspectives. The perspectives included are from a journal editor, a university admissions researcher, two leaders from admissions testing organizations (one from the USA and one from Chile), and a validity researcher-practitioner who suggests a change in perspective on how we validate admissions tests. Each presenter will share their opinions and expertise, and the presentations will be followed by remarks from an early career discussant with experience in validity theory, test validation, and issues affecting minoritized students. Following the discussion, time will be provided for Q&A with the audience.

Session Organizer:

Stephen Sireci, University of Massachusetts Amherst

Moderator:

Monica Silva, Pontificia Universidad Católica de Chile

Presenters:

Deborah Harris, University of Iowa
Adele Brumfield, University of Michigan
Diane Henderson, ACT
Alvaro J. Arce, Pearson
Leonor Varas, Demre

Discussant:

Joseph A. Rios, University of Minnesota

113. Machine Learning Topics

Paper Session
9:45 to 11:15 am PT
Westin San Diego Gaslamp: California Ballroom B

Participants:

Interactive Score Reporting Dashboard Improvement through NLP Analysis of Focus Group Feedback

Merve Sarac, UW-Madison; Rich Feinberg, National Board of Medical Examiners; Francis O'Donnell, National Board of Medical Examiners; Thai Quang Ong, National Board of Medical Examiners

This research study investigated focus group feedback on an interactive score reporting dashboard presenting results for multiple related assessments. Data were analyzed using two NLP methodologies: topic modeling and sentiment analysis. Results validated qualitative findings, provided additional insight on key themes, and clarified next steps for expanding the dashboard's utility.

Assessing human-computer rater agreement with Bayesian equivalence testing

Sydne McCluskey, CUNY Graduate Center; Liuhan Cai, Cognia; Xi Wang, Cognia; Louis Roussos, Cognia

As automated scoring is increasingly used operationally, understanding how it compares to human scoring is essential. Agreement indices are commonly used for this purpose, but they can miss systematic differences between raters. This paper applies Bayesian equivalence testing to a measurement invariance framework to compare human and computer raters.

Improving equity in AI-based learner modeling using aggregated demographic data

David King; Kamil Akhuseyinoglu, University of Pittsburgh; M Hassan Musturi, Edmentum; Ziwei Zhou, Edmentum

This paper examines opportunities and challenges of implementing deep knowledge tracing for measuring learner skills in a production adaptive instructional system, with a focus on evaluating fairness across learners by Title 1 status, metro classification, gender, and ethnicity. Sampling techniques for obtaining representative training samples are explored for improving equity.

Teachers' Perceptions of The Validity of an Automated Writing Evaluation System

Yue Huang, University of Delaware; Andrew Potter, University of Delaware; Joshua Wilson, University of Delaware

We explored teachers' perceptions of the validity of feedback from an automated writing evaluation (AWE) system. Focus groups were conducted among upper-elementary teachers (N = 13). Findings revealed that teachers perceived of AWE feedback to be generally reliable and valid, but feedback may not be appropriate for all students.

The Effect of Word Vector Representation and Linguistic Features on the Accuracy of Automated Essay Scoring Systems Using Neural Networks

Tahereh Firoozi; Ali Naeim Abadi, University of Alberta; Carrie Demmans Epp, Department of Computing Science, University of Alberta; Okan Bulut, University of Alberta; Denilson Barbosa, Department of Computing Science, University of Alberta

This study examined the effect of Word2Vec and Glove embedding techniques as well as linguistic features on the accuracy of the automatic essay scoring (AES) system using an LSTM model. The results showed that word embedding techniques significantly improved the accuracy of the model (QWK= 0.79). However, the linguistic features did not contribute to the accuracy of the LSTM model.

Discussant:

AJ Alvero, Stanford

114. eBoard Session

Electronic Board Session

9:45 to 11:15 am PT

Westin San Diego Gaslamp: San Diego Ballroom

Participants:

Accuracy in Invariance Detection with Multilevel Models with Three Estimators

Holmes Finch, Ball State University; Brian French, Washington State University; Thao Vo, Washington State University; Cihan Demir, Washington State University

Applied and simulation studies document model convergence and accuracy issues in differential item functioning detection with multilevel models. This simulation study investigated such issues for invariance detection with multilevel logistic regression models with predictors at level 2. Preliminary results suggest the Bayes estimator performed best compared to alternatives.

Involving Cognitive Diagnostic Models in Test Development and Use: A Systematic Review

Hongli Li, Georgia State University; Charles Vincent Hunter, Clayton County Public Schools; Ren Liu, University Of California, Merced

Many times, cognitive diagnostic models (CDMs) are retrofitted to preexisting non-diagnostic tests, which limits the full potential of CDMs. In this study, we conducted a systematic review to describe the current picture of involving CDMs in the test development stage and consequently using the tests for diagnostic purposes.

A simulation study applying Psychometric Network Analysis with Clinical Measurement Data

Hsin-Ro Wei, Riverside Insights; Michael Custer, Riverside Insights; Edward Wolfe, Riverside Insights; David Dailey, Dailey data group

This study utilizes a clinical dataset and psychometric network analysis (PNA) to study PNA as a methodology to confirm an underlying data structure. This study investigates how varying levels of sample size, proportion of missing subtests, and proportions of missingness among test takers impact psychometric network analysis accuracy and stability.

The Effect of Test Length on Preknowledge Detection Using Ls and Rs

Hyun Joo Jung

Sinharay (2017) suggested the signed likelihood ratio statistic and the signed score statistic to detect item preknowledge. He mentioned different test lengths may possibly affect the performance of the statistics. In this study, we investigate their performance across various conditions – test length, IRT models, and sample size.

Current State and Future Directions of Performance-based Clinical Competence Assessments

Igor Himelfarb; Brooke Houck, NBEO; Nai-En Tang, National Board of Chiropractic Examiners; Andrew Gow, National Board of Chiropractic Examiners

On January 26, 2021, the USMLE cancelled Step 2 Clinical Skills exam, which prompted an evaluation of performance-based clinical competence assessment throughout the healthcare testing organization. This study is a critical look into the assessment of clinical skills in the fields of Chiropractic and Optometry.

Testing Normal Distribution Assumptions of the Cutscore Operating Function

Jesse R. Pace

Grabovsky and Wainer's (G&W; 2017) method to estimate classification error and optimal cutscore location assumes a normal distribution for examinee true scores. The present research explores the performance of G&W estimates under increasingly non-normal Monte Carlo simulations (i.e., skewed, kurtotic, and bimodal). Implications for standard setting are discussed.

Simultaneous Linking for Maintaining the Vertical Scale of a Two-Level Test

Jiyun Zu, Educational Testing Service; Lixiong Gu, Educational Testing Service

We study different linking designs for maintaining vertical scales using real data from a two-level English language test. We hypothesize that it is beneficial to embed both within-level and between-level common items and use simultaneous linking method. Between-level common item parameter estimates, linking coefficients, and new form conversions are compared.

Comparison of pre- and post-equating methods using empirical data in post-pandemic environment

Joanna Tomkowicz, Data Recognition Corporation; Dong-In Kim, DATA Recognition Corporation; Ping Wan, Data Recognition Corporation

This study evaluated stability of item parameters and student scores, using the pre-equated (pre-pandemic) parameters from Spring 2019 and post-equated (post-pandemic) parameters from Spring 2021 in three approaches to anchor treatment: re-estimating anchor parameters, holding c-parameter fixed for multiple-choice anchors, and holding all anchor parameters fixed to their pre-pandemic values.

Modeling Response Processes in Early Literacy Measure: An Explanatory IRT Approach

Jose R. Palma, The University of Texas at Austin

This study looked at the utility of explanatory IRT in detecting the impact of response processes on the functioning of an early literacy measure of phonological awareness. We found significant effects of person and item characteristics. Impact of these variables on the measure is further examined via a simulation study.

Standard Errors for Matched Samples

Junhui Liu, Educational Testing Service; Ru Lu, Educational Testing Service

Statistical procedures have been used to match samples of students to compare ability differences across groups. Little attention was paid to the standard error of modeled effects after matching. This study intends to investigate the proper ways to estimate standard error for significance tests in matched or weighted samples.

Using the IRTree Model to Investigate Moderation Effects in Response Processes

Justin L. Kern, University of Illinois at Urbana-Champaign; Auburn Jimenez

IRTree models are IRT models that assume a tree structure for response processes. They have been shown to be useful in investigating response styles, omitted responses, and ambiguous response options. In the current project, we show that IRTree models can be used to investigate within-item moderation effects.

Evaluation of theory-driven, two-domain based routing algorithm in language testing

Kyoungwon Lee Bishop, WIDA at University of Wisconsin Madison; Sakine Gocer Sahin, WIDA at UW-Madison

This study evaluates and redesigns a two-domain integrated routing algorithm in a multistage adaptive test (MST) of an academic English language proficiency assessment (MODEL) for English language learners (ELLs).

Cross-Cultural Image-based Assessments Lisa Keller, University of Massachusetts;

Kevin O'Rourke, University of Massachusetts Amherst

It is often said that a picture is worth a thousand words. This study investigates the feasibility of using image-based assessments across different cultures. The results are compared with traditional text-based assessment. The findings suggest that there is promise in using image-based assessments cross-culturally. Image-based assessments have appeal for special populations, cross-cultural assessment, and to increase enjoyment of, and engagement with assessment.

A Usability Analysis of Available R Packages for Text Clustering Methods

Magdalen Beiting-Parrish, CUNY Graduate Center; Christopher Runyon, NBME

Text mining has exploded in psychometric research recently. One extremely useful method is text clustering, which can help the researcher make sense of latent topics/themes within a collection of texts. This research evaluates all the available R text clustering/visualization packages available on the CRAN for usability for novice researchers.

The Performance of the FIML estimator with Missing Binary Predictors

Michael Wayne Harris, University Of South Carolina; Dexin Shi, University of South Carolina; Amanda Fairchild, University of South Carolina

We investigated the robustness of the performance of the full information maximum likelihood (FIML) estimator with missing binary predictors, when the multivariate normality assumption is violated.

Diagnostic Classification Modeling of English Language Vocabulary Knowledge for International Test Takers

Mingying Zheng, University of Iowa; Geoff LaFlair, Duolingo

This study is to explore the usefulness of diagnostic classification modeling (DCMs) in a high-stakes English language test of vocabulary knowledge. The results are reported regarding the diagnostic quality of items, attribute mastery profiles, attribute relationships, and diagnostic feedback to test takers at individual level and at group level.

Studying Response Time and Performance with Students' Online Research and Comprehension Ability

Ya Mo; Brian Habing, National Institute of Statistical Sciences; Nell Sedransk, National Institute of Statistical Sciences; Alexi Albert, National Institute of Statstical Sciences

The similarity of performance-based assessment (PBA) and multiple-choice assessment (MC) when assessing students' online research and comprehension ability is an open question. This study advances previous research by accounting for students' responses and response times to two different item types: traditional multiple choice and free response using a synthetic internet.

A Comparison between CAT and Decision Tree with Computerized Adaptive Training-Testing Program

Yiling Cheng, Kaohsiung Medical University; Mark Reckase, Psychometric Solutions

The goals of the study are two folds: performance comparisons between computerized adaptive testing (CAT) and decision tree methods on multidimensional datasets are conducted with a simulation study. A realistic, tutorial example using Concerto to create such a multidimensional training-testing program will also be presented

Impact of Embedded Field Testing on MCAT® Examinee Scores

Ying Jin, Association of American Medical Colleges; Charles Fisk, Associate of American Medical Colleges; Marc Kroopnick, Association of American Medical Colleges

The study examined the impact of embedded field testing on MCAT® examinee scores. The findings suggest that, while the difficulty of field test items varied across versions of an operational form, such differences did not impact examinee performance on the scored portion of the form.

Simultaneous linking for Improving Scale Stability

Ying Lu, College Board; Judit Antal, College Board

This study compares the performance of simultaneous linking to that of the Stocking and Lord procedure and IRT fixed anchor calibration in maintaining scale stability. This study also examines how simultaneous linking performs under moderate item parameter drift and explores how to effectively monitor scale drift under simultaneous linking.

115. Broken Systems of Assessment: Addressing Fairness and Equity Challenges in Special Populations

Organized Discussion
9:45 to 11:15 am PT
Westin San Diego Gaslamp: Del Mar

Equity and fairness are not new challenges in the assessment world. Issues of bias, inequity, and unfair testing practices have plagued our educational system for decades, and continue to do so. Despite a great deal of progress, specific equity and fairness challenges continue to exist in our education and assessment systems. Inequities that were well known pre-pandemic are now exacerbated, setting us even further behind. Special populations, including students with special needs, English language learners, minority racial and ethnic groups, and low-income students, have been most impacted by these issues with regard to general education and assessment. Educators, measurement experts, and testing professionals have attempted to understand these populations in order to create more refined tests, define needed support systems, and establish policies that meet their specific needs in a fair and equitable manner. In this session, our diverse panelists will share their various definitions and interpretations of terms such as "special" and "fairness" and propose collaborative efforts that help bridge existing gaps between them. They will engage in a healthy debate around the appropriate/inappropriate uses of tests and will close with innovations/recommendations that get us closer to an assessment that results in fair and equitable measures for all students.

Session Organizer:
Dubravka Svetina Valdivia, Indiana University

Presenters:
Elda Garcia, National Association of Testing Professionals
Gregory Cizek, University Of North Carolina
Melissa L. Gholson, Educational Testing Service (ETS)
Pohai Kukea Shultz, University of Hawaii at Manoa

116. Software Demonstrations

Demonstration Session
9:45 to 11:15 am PT
Westin San Diego Gaslamp: La Jolla

Participants:

Interactive NAEP Results Display, an R Shiny Tool

Michael Lee, AIR; Emmanuel Sikali

This demonstration covers the interactive NAEP Results Display, an R Shiny application used to analyze and visualize the NAEP performance of critical demographic groups over year in various subjects. The tool's intuitive interface exploring student results makes self-guided knowledge acquisition accessible to policy makers, researchers, and the general public.

Is Everyone Simulating ShadowCATs Without Me? (and Other Form Construction Concerns)

Samuel Haring, ACT; Bingnan Jiang, ACT

RSCAT is an R package provides a shadow-test simulator that uses a free copy of FICO Xpress for the modeling. This demonstration will show how to obtain the simulator, run a simulation, and discuss potential novel uses of the simulation results such as rudimentary form construction and MST panel building.

What Lies Beyond? Process Data iSmart Tool

Ruhan Circi, American Institutes for Research; Juanita Hicks, American Institutes for Research; Tiago A. Caliço, American Institutes for Research; Emmanuel Sikali

Since NAEP's transition from paper- to digitally based assessments, data on students' interactions with each item became available. These data are called process data. AIR developed an item information platform that extends traditional item characteristic information with process data, to make results about process data accessible and interpretable to educators.

maat: An R Package for Multiple Administrations Adaptive Testing

Seung W. Choi, University of Texas At Austin; Christina Schneider, Cambium Assessment

This session is a demonstration of the maat package which is proof-of-concept for a through-year, or hybrid interim-summative, computer adaptive assessment system. The configurable test design, technical documentation, and evidence of functionality will be showcased along with the test purposes the assessment design was created to meet.

117. Measuring the impact of COVID: Methodological challenges in understanding unfinished learning

Organized Discussion
9:45 to 11:15 am PT
Westin San Diego Gaslamp: Santa Fe

Following the cancellation of summative assessments in spring 2020 due to COVID, many states and districts looked to interim assessment results to provide insight on learning and the impact of schooling disruptions. As the pandemic has continued, many interim assessment providers have taken an additional role to inform larger educational policy discussions by publishing research on national trends in student performance and growth, as well as diving into questions about equity and differential impact for various student and school demographic groups. While interim assessment results can provide timely insights into how our nation's students and schools are faring and help fill the void left by continued disruptions in summative testing, the interpretations of these assessment data are impacted by methodological choices and data quality. For example, researchers have had to wrestle with new threats to data quality including those from unproctored remote testing, missing students, and differential demographic representation in remote, in-school, and hybrid testing and learning environments. This session will share a behind the scenes look at how researchers at four different interim assessment organizations navigated these challenges and how these choices influenced their reported (or unreported) findings.

Session Organizer:
Laurie Davis, Curriculum Associates

Moderator:
Andrew Ho, Harvard Graduate School of Education

Presenters:
Megan Kuhfeld, NWEA
Katie McClarty, Renaissance
Logan Rome, Curriculum Associates
Audra Kosh, Edmentum, Inc.

118. IRT Equating and Linking

Paper Session
9:45 to 11:15 am PT
Westin San Diego Gaslamp: Plaza

Participants:

Approaches to Linking Different Item Response Theory Score Scales

Jaime Malatesta, Graduate Management Admission Council; Mengyao Zhang, National Conference Of Bar Examiners; Mark R Connally, National Conference of Bar Examiners; Won-Chan Lee, University of Iowa

As testing programs evolve, practical challenges can make it desirable to adopt a different item response theory model. This study evaluates the application of three linking methods for a situation where the 3PL model is replaced with the Rasch model, but it is desirable to maintain the original scale.

Effects of Calibration Approach and Ability Distribution on IRT Observed-Score Equating Results

Hyung Jin Kim, University of Iowa; Won-Chan Lee, University Of Iowa; Tianyou Wang, University of Iowa; Robert Brennan, University of Iowa

For IRT observed-score equating, results can depend on choices for two factors: (1) an item calibration approach, and (2) an ability distribution for constructing fitted score distributions. Using real data, the study considers various options for the two factors and compares results to provide practical insights about IRT observed-score equating.

Evaluating Several Variants of Simple-Structure MIRT Equating

Stella Kim, University of North Carolina at Charlotte; Won-Chan Lee, University Of Iowa

The primary purposes of this study are to propose new true-score and observed-score equating methods under simple-structure multidimensional IRT (SS-MIRT) and compare them with existing SS-MIRT equating methods. Results from the proposed methods are evaluated with respect to equating accuracy under several simulation conditions.

Finding Stability: Comparing Methods for Detecting Unstable Item Parameters in IRT Equating

Jeffrey Steedle, ACT, Inc.

Instability in common item parameters can introduce error into IRT scale transformations. This study compared five methods of identifying items with significant parameter drift. Rather than simulating data, this study took advantage of real-world data from random groups equipercentile equating to learn something about common item nonequivalent groups equating.

Accuracy of Item Parameter Linking Accounting for Uncertainty in Parameter Estimates

Jessica Plourde, Fordham University; Leah Feuerstahler, Fordham University

This project compares three IRT linking methods – the Haebara method, concurrent calibration, and a recently developed method that accounts for errors in item parameter estimation – in correctly specified and misspecified models. Results are mixed, but suggest that accounting for item errors provides little advantage over existing methods.

Discussant:

Anthony Albano, University of California, Davis

119. Improving the Accuracy of Aggregate Growth Measures: Statistical Methods and Practical Challenges

Coordinated Paper Session
9:45 to 11:15 am PT
Westin San Diego Gaslamp: California Ballroom C

Currently, 49 states in the U.S. measure student achievement growth as part of their statewide student-testing programs. These measures use the average of individual students' growth for all the students in a district, school, or educator's class. Although this approach may appear to be a practical and straightforward, simple averages of growth measures for few students can have substantial year-to-year fluctuations, limiting the value of the average of student growth for informing decision making. In response to high year-to-year fluctuations of the simple average growth scores, in California, ETS researchers applied the Empirical Best Linear Prediction (EBLP) methodology to aggregate growth. EBLP can improve the accuracy of the growth measures and in turn year-to-year stability This session highlights this methodology, and implications for the adoption of any new growth model or methodology in a state system. Specifically, the session includes: 1. a review of the EBLP methodology for estimating aggregate growth scores, 2. an investigation the EBLP methodology when applied to two states' data, 3. a walkthrough of the communication outreach efforts with key stakeholders, including lessons learned and modifications to plans in response to stakeholder feedback, and 4. discussion of the papers by Damian Betebenner, a national growth expert.

Session Organizer:

Daniel McCaffrey, Educational Testing Service

Participants:

Improving Accuracy and Stability of Aggregate Student Growth Measures Using Empirical Best Linear Prediction

Daniel McCaffrey, Educational Testing Service; J.R. Lockwood, Duolingo; Katherine Furgol Castellano, Educational Testing Service

Applications of the Empirical Best Linear Prediction to Aggregate Growth Measures in Two States
Katherine Furgol Castellano, Educational Testing Service; Daniel McCaffrey, Educational Testing Service; J.R. Lockwood, Duolingo

Adopting a Growth Measure in California
Kimberly Mundhenk, California Department of Education; Katrina Gonzalez, California Department of Education

Discussant:
Damian Betebenner, National Center for the Improvement of Educational Assessment

120. What Does A Socially Responsible Future Look Like For Admissions Testing?

Coordinated Paper Session
11:30 to 1:00 pm PT
Westin San Diego Gaslamp: California Ballroom A

In recent years, we have seen a heightening in public attention toward systemic racial inequalities and other socioeconomic divides. This, combined with the advent of the COVID-19 pandemic, has precipitated a flood of states and institutions abandoning required college admissions tests in favor of test-optional policies, totalling more than two-thirds of U.S. four-year colleges and universities for both Fall 2021 and 2022 admissions (FairTest, 2020; FairTest, 2021). At this critical juncture, the admissions testing community faces widespread public distrust and dissatisfaction, and needs to chart a new path forward. In the spring of 2021, Educational Measurement: Issues and Practice editor Deborah Harris invited the measurement community to engage with a focal article by Koljatic, Silva, and Sirici on "College Admissions Tests and Social Responsibility", with the aim of eliciting discussion and debate around this important topic. In this symposium, we bring together the authors of several of these commentaries to discuss what a socially responsible future for admissions testing might look like. Contributing authors from both industry and academia will put forth their various visions, next steps, and research agendas for the measurement community in brief presentations, allowing plentiful time for cross-discussion and audience questions.

Session Organizer:
Emma M. Klugman, Harvard Graduate School of Education

Moderator:
William Harris, Assoc. Of Test Publishers

Participants:
The Questions We Should Be Asking About Socially Responsible College Admission Testing
Emma M. Klugman, Harvard Graduate School of Education; Lily An, Harvard Graduate School of Education; Zach Himmelsbach; Sophie Litschwartz; Tara P. Nicola, Harvard Graduate School of Education

Evolution of Equity Perspectives on Higher Education Admissions Testing: A Call for Increased Critical Consciousness
Susan Lyons, Lyons Assessment Consulting; Fiona Hinds, Cognia, Inc.; John Poggio, University of Kansas

Reviving the Messenger: A Response to Koljatic et al. (2021)
Krista Mattern; Ty Cruce; Dianne Henderson, ACT; Tina Gridiron, ACT, Inc; Alex Casillas, ACT, Inc; Melinda Ann Taylor, ACT

Social Responsibility in College Admissions Requires a Reimagining of Standardized Testing
Anthony Albano, University Of California, Davis

Achieving Educational Equity Requires a Communal Effort
Michael E. Walker, Educational Testing Service

Discussant:
Stephen G Sireci, University of Massachusetts, Amherst

121. Fit Statistics

Paper Session
11:30 to 1:00 pm PT
Westin San Diego Gaslamp: California Ballroom B

Participants:
Development and Evaluation of a Composite Item-Fit Statistic for Diagnostic Classification Models
Jennifer Kobrin, ATLAS: University of Kansas; William Jacob Thompson, University of Kansas; Wenhao Wang; Jeffrey Hoover, University of Kansas
We examined a new composite item-fit statistic for a diagnostic assessment system. The composite synthesizes multiple item-fit statistics as an estimate of item quality. We investigated the utility of the composite by comparing test development staff item ratings with categories and severity levels of the composite.

Effects of Item Misfit on Proficiency Estimates Under the Rasch Model

Chunyan Liu, National Board of Medical Examiners; Peter Baldwin, National Board of Medical Examiners; Raja G Subhiyah, National Board of Medical Examiners

When IRT parameter estimates are used to make inferences about examinee performance, assessment of model-data fit is an important consideration. The findings of this simulation study suggest that item misfit due to violating the equal discrimination assumption under Rasch model can have deleterious effects on ability estimation and examinee classification.

Evaluating Goodness-of-Fit at the Testlet Level in the Rasch Testlet Model

Dandan Liao, Cambium Assessment, Inc.; Frank Rijmen, Cambium Assessment, Inc

The present study generalizes the existing goodness-of-fit statistics for the Rasch model to assess model fit for items sharing the same stimulus in the Rasch testlet model. Type I error rate and power of the proposed statistics were evaluated via simulation studies.

Impact of item misfit on Winsteps calibrations

Wenli Ouyang, National Board of Medical Examiners; Chunyan Liu, National Board of Medical Examiners; Raja G Subhiyah, National Board of Medical Examiners

The practical consequences of equal discrimination assumption violations under the Rasch model were investigated based on simulated data. The factors manipulated include the proportion and magnitude of misfitting items. The results suggested that both item parameter estimates and Rasch fit statistics were negatively impacted for both misfitting and well-fitting items.

Person-Fit Statistics Complement Traditional Methods of Identifying Careless Responders in Surveys

Eli Andrew Jones, The University of Memphis; Stefanie A. Wind, University of Alabama; Chia-Lin Tsai; Yuan Ge, University of Alabama

Careless responding complicates the interpretation and use of results from self-report measurement procedures. We examined the alignment between traditional methods for identifying carelessness in survey research (e.g., observational and outlier methods) with person fit analysis from IRT. Results indicate moderate-to-strong correlations and suggest that these statistics provide complimentary information.

Discussant:

Leah Feuerstahler, Fordham University

122. GSIC eBoard Session

Graduate Electronic Board Session

11:30 to 1:00 pm PT

Westin San Diego Gaslamp: San Diego Ballroom

Participants:

Anchor Selection Strategies Using Multiple-Group Categorical CFA

Haeju Lee, University of North Carolina Greenboro; Kyung Yong Kim, University of North Carolina Greenboro

In this study, we investigate the effective anchor selection procedure that is applicable to categorical confirmatory factor analysis (CCFA) models for testing measurement invariance. CFA based-likelihood ratio test (LRT) with backward approach will be used as the anchor selection procedure.

A recommendation for the degree of variance to apply measurement alignment

SEOHEE PARK, American Board of Internal Medicine; Seongeun Kim, University of North Carolina at Greensboro; Hyun Suk Ryoo, Gerorge Town University; Ji Hoon Ryoo, Yonsei University

Measurement alignment is recently introduced and gradually used in many different fields. However, an assumption of measurement alignment, a majority of items should be invariant, is not specified enough. To provide a better perception of the extent of invariance, this study explores various levels of variance through a simulation study.

Direct and indirect evidence of ICT treatment types for second language learning: a network meta-analysis

Songtao Wang, OISE/University of Toronto

This multi-arms network meta-analysis study examined the effects of different technological treatments for second language (L2) learning. 29 experimental studies were included which yielded 50 effect sizes. Results showed ICT treatments using textual prompts with medium to large effect sizes while those using graphic prompts showed lower effect sizes. The within-designs heterogeneity was statistically significant while no between-designs inconsistency was found. Practical implications on ICT treatments, network meta-analysis and measurement validity were discussed.

Effect of Ability Distribution on IRT Fitted Score Distribution

Huan Liu, The University of Iowa; Won-Chan Lee, University of Iowa; Hyung Jin Kim, University of Iowa

This study aims to investigate empirically the effect of using different ability distributions on IRT fitted score distributions using a large number of real datasets. Preliminary results showed that the posterior distribution produces closer fitted score distribution to the observed score distribution than the assumed standard normal distribution.

Effects of Varying Item-Pool Size in the Randomesque Item Exposure Mechanism

Ted Daisher, Curriculum Associates

Using simulations of an adaptive K-12 test in mathematics and reading, this study compares candidate item-pool sizes (1, 3, 5, 7, and 8 items) in the randomesque method of controlling for item exposure by how they affect test length, item overlap, item exposure, pool utilization, and item repeat.

Enhancing the Identification of Test-Taking Disengagement with Hint-Taking Behavior

Guher Gorgun, University of Alberta; Seyma N. Yildirim-Erbasli, University of Alberta

In this study, we investigated the relationship between hint-taking behavior and test-taking engagement. We found that disengaged students tend to exploit hints and use many hints in a short amount of time compared to engaged students. Modeling hint-taking behavior might be feasible for the detection of test-taking disengagement.

Evaluating Sensitivity and Specificity of PPMC method

Mingjia Ma, The University of Iowa

This paper tries to investigate the sensitivity and specificity of PPMC for two IRT model discrepancy measures from the perspective of prior selection in detecting model misfit and different grouping criterion in discrepancy measures.

Identifying Latent Classes with Explanatory IRT models for Response and Response Times

Clifford Erhardt Hauenstein; Susan Embretson, Georgia Institute of Technology

A mixture explanatory item response model for both response and response time is proposed. The model follows the tradition of IRT models that evaluate process validity by relating item features to item parameters, and identifies latent clusters based on distinct response processes when response and response time data is collected.

Identifying Response Time Thresholds for Solution Behaviors

Yi Yang, Columbia University; Young Yee Kim, American Institute for Research Xiaying Zheng, American Institutes For Research; Xiaoying Feng, American Institutes for Research

Differentiating between effortful behavior (solution behavior) and non-effortful behavior (skipping, rapid guessing, etc.) can contribute to ensuring the validity of achievement test scores. This study explores the effective approach to identifying meaningful interactions for both item responses and item nonresponses by applying multiple threshold-setting methods to NAEP data.

Investigating the Impact of the Anchor Item Preknowledge on Test Equating

Onur Demirkaya, University Of Illinois at Urbana-Champaign; Ummugul Bezirhan, Boston College

Anchor items are used in multiple test forms under the nonequivalent anchor test (NEAT) design. If item preknowledge occurs on one of the forms, the equating procedure may produce dubious results. This study examines the impact of anchor item preknowledge under the NEAT design through a simulation study.

IRT Observed-Score Equating for Rater-Mediated Assessment using a Hierarchical Rater Model

Tong Wu, University of North Carolina Charlotte; Stella Kim, University of North Carolina at Charlotte; Carl Westine, University of North Carolina at Charlotte; Michelle Boyer, Data Recognition Corporation

The purpose of this study to propose an IRT observed-score equating method under the Hierarchical Rater Model (HRM) framework to account for rater errors in equating. Equating accuracy of the proposed equating method and the traditional IRT observed-score equating method is compared and evaluated under various simulation conditions.

Issues of Validity Regarding Assessment Practices In Universities In Ghana During Covid-19

Frank Quansah, University of Cape Coast

The emergence of the covid-19 pandemic led to the upsurge of online (remote) modes of teaching, assessment, and learning in universities across the globe. In an attempt to explore the assessment practices adopted, this study found that there are issues concerning the validity of the assessments conducted during the period.

Item Parameter Estimation for Multidimensional Graded Response Model Under Complex Structure

Olasunkanmi Kehinde, Washington State University; Shenghai Dai, Washington State University; Brian French, Washington State University

The current study examines item parameter recovery in partial compensatory multidimensional graded response model (PC-MGRM) under complex structure with rating-scale items. A simulation study was performed to investigate factors that might influence the precision of item parameter estimations, including sample size, intercorrelation between the dimensions, test lengths, and dimensional structures.

Predictors Of Adoption Of Computer-Based Testing In Ghana

Francis Ankomah, University of Cape Coast

The quest to embrace ICT, coupled the measures to curb covid-19, call for the practice of computer-based assessment in Ghana. This study seeks to identify the predictors of adoption of computer-based testing from the perspective of teachers and students in Ghana.

Response Styles in Multiscale Assessment: A Method of Plausible Value Distribution

Zebing Wu

A plausible value (PV) method is proposed to identify response styles across multiple subscales and compared to MIRT and IRTree methods. A simulation and two real datasets are analyzed to investigate its effectiveness. Combining with other methods, the PV method may improve measurement precision for substantive traits.

Sequential Pattern Mining to Understand Time Management Strategies in Online Courses

Seyma N. Yildirim-Erbasli, University of Alberta; Guher Gorgun, University of Alberta; Yizhu Gao; Okan Bulut, University of Alberta

This study explores students' time management strategies in an online learning setting where students could study at their own pace. The analysis of trace data of learning behaviors showed that different time management strategies exist across student clusters, which are also related to students' time management tactics.

Teacher Perceptions of Equity and Justice in Classroom Assessment: An Exploratory Study

Janine Jackson, Morgan State University

In this study, secondary teachers' abilities to translate the principles and expectations of culturally responsive, culturally relevant, culturally sustaining, antiracist, and justice-oriented approaches to pedagogy into classroom assessment practices are explored.

Validity of People in my Life- Spanish version: Multidimensional Graded Response Study

Sandra Liliana Camargo Salamanca, Purdue University

This study aims to analyze the PIML internal structure through a multidimensional graded response model to improve its scoring and interpretation. 2273 children between 8 and 14 years of age from Colombia participated in this study. A more precise scoring system and score interpretation of PIML are proposed.

123. Students with Disabilities

Paper Session
11:30 to 1:00 pm PT
Westin San Diego Gaslamp: Del Mar

Participants:

A Tale of Two Disabilities: Findings from NAEP Response and Process Data

Xin Wei, SRI International; Jihong Zhang, University of Iowa; Susu Zhang, University of Illinois at Urbana-Champaign

Using 8th-grade NAEP performance and process data, our differential item functioning and differential response time analysis based on a conditional permutation test revealed different performance and response time patterns for students with autism spectrum disorder and students with learning disabilities compared to their typically developing peers.

Parent and Youth Expectations on Post-school Outcomes for Youth with Disabilities

Yi-Chen Wu, University of Minnesota/NCEO; Martha Thurlow, National Center On Educational Outcomes; David Johnson, University of Minnesota

This study examined parent and youth expectations on post-school outcomes for students with IEPs using the National Longitudinal Transition Study–2012 (NLTS 2012) dataset and the agreement between parent and youth expectations was examined. Further, these expectations were examined between NLTS 2012 and NLTS2 data.

Subgroups of Mathematical Learning Disability: A New Classification Method based on Cognitive Diagnostic Models and Their Cognitive-linguistic Correlates

Xiangzi Ouyang, The university of Hong Kong; Xiao Zhang, The university of Hong Kong; Qiusi Zhang, Purdue University

The present study, using Cognitive Diagnostic Modelling (CDM), identified six most prevalent MLD subgroups in a sample of 204 MLD children. The reliability and validity of the classification were evaluated by comparing with the control group (children with low achievement in mathematics). We also examined their cognitive-linguistic correlates.

Usage of Accessibility and Universal Design Features in Digitally Based Assessment and Student Performance

Juanita Hicks, AIR; Ruhan Circi, American Institutes for Research; Burhan Ogut, American Institutes for Research; Darrick Yee, American Institutes for Research; Michelle Yin, Northwestern University

Research is scarce on whether and how students with disabilities use accessibility features in digitally based assessments. Using process data collected in the 2017 NAEP Grade 8 mathematics assessment, this study provides a first look at the usage patterns of accessibility features and how they correlate with student academic performances.

Discussant:

Richard Patz, University Of California Berkeley

124. Automatic Item Generation: Research & Applications

Coordinated Paper Session
11:30 to 1:00 pm PT
Westin San Diego Gaslamp: La Jolla

This coordinated session explores research and applications using automatic item generation throughout the full cycle of assessment, which include item generation and banking, development and evaluation, as well as delivery and administration. The first session describes and illustrates methods that use metadata to organize generated items in a bank as well as identify enemy items, analyze a bank, enhance their test security practices, protect against fraudulent test administrations, and personalize student learning. The second session examines the quality of generated items for low-and high-stakes assessments. Studies of simulated and real candidate data from national professional credentialing programs will be summarized, including calibration and prediction of psychometric characteristics of item models and survival rates of AIG vs. traditional items. The last session presents an alternate use of AIG item models or templates, the templates themselves serving as items on tests instead of generating other items to do so. The value of and theoretical support for such unique testing will be discussed as will results from a 3-year case study. The panelists represent diverse perspectives from both academic and industry assessment professionals with their current research interest and latest research results.

Session Organizer:

Xin Li, PSI Services

Participants:

Automatic Item Generation and the Challenge of Item Banking
Hollis Lai, University of Alberta; Mark Gierl, University of Alberta

Take AIG to the Next Level for both High- and Low-stakes Assessment Development
Xin Li, PSI Services; John Weiner, PSI Services, LLC

Item Models for AIG Used Directly in Test Administration
David Foster, Caveon Test Security

Discussant:
Suzanne Lane, University Of Pittsburgh

125. AERA Division D and NCME: Building on our Synergies and Singularities

Organized Discussion
11:30 to 1:00 pm PT
Westin San Diego Gaslamp: Santa Fe

AERA Division D focuses on measurement, psychometrics, and assessment; statistical theory and quantitative methodologies; qualitative methodologies; and multiple and mixed methodologies as applied to educational research. The National Council on Measurement in Education (NCME) is a professional organization for individuals involved in assessment, evaluation, testing, and other aspects of educational measurement. This joint session has organized a panel of speakers from both Division D and NCME to discuss the similarities and differences of two organizations and how we could work together to better serve the broader community.

Chair:
Brian C Leventhal, James Madison University

Moderator:
Mary Pitoniak, ETS

Presenters:
Wayne J. Camara, LSAC
Linda Cook, ETS
Kathryn Nicole Thompson, James Madison University
Stephanie Shelton, The University of Alabama

126. Classroom Assessment and Data Literacy

Paper Session
11:30 to 1:00 pm PT
Westin San Diego Gaslamp: Plaza

Participants:

Adaptation of Inventory of Classroom Assessment and Evaluation Approaches to Turkish Culture
Yesim Ozer Ozkan, Gaziantep University
The aim of this research is to adapt the Inventory of Classroom Assessment and Evaluation Approaches in Turkish, developed by DeLuca, Valiquette, Coombs, & LaPointe-McEwan (2018). This inventory provides teachers with a personal assessment profile that defines classroom assessment approaches.

Classroom Assessment Experiences of College Students during COVID-19
Teresa Ober, University of Notre Dame; Maxwell Hong; Kathleen Morse, University of Notre Dame; Ying Cheng, University of Notre Dame
Classroom assessment experiences of U.S. college students (N=992) are summarized, providing insight into the common types of modifications that were made and are continuing to evaluate student learning. The results provide a deeper understanding of the longer-term impact of COVID-19 on today's college students and stakeholders of assessment outcomes.

Classroom Assessment Standards-Based Grading
Corrie Rebecca Klinger, University of Waikato
The Classroom Assessment Standards (Klinger, et al., 2015) were the criteria for mastery in this educational measurement study of a standards-based grading approach teaching preservice teachers to evaluate student learning. From 2014 through 2018, the study was conducted. The 109 participants were becoming elementary and middle school teachers.

Teachers' Classroom Assessment Practices in the U.S.: Evidence from PISA 2018
Roti Chakraborty, Georgia State University; Phylicia Thompson, Georgia State University; Hongli Li, Georgia State University
This study used PISA 2018 data to understand the classroom assessment practice in the U.S. We explored what kinds of classroom assessment practices are used in the U.S and what teacher-level and school-level characteristics explain the variation of teachers' classroom assessment practices based on 1,812 teachers within 158 schools.

Investigating the impact of cognitive biases on data literacy performance

Alina Lutsyk; Jacqueline P. Leighton, University of Alberta; Ying Cui, University of Alberta; Fu Chen; Maria Cutumisu, University of Alberta

One hundred undergraduate students took part in an innovative digital data literacy assessment. The objective of this study was to identify the relationship between data literacy skills and three systematic thinking biases (e.g., confirmation, belief, ignoring P(D/~H) biases) that affect evaluation of data and decision-making with the use of data.

Discussant:

Bozhidar M. Bashkov, IXL Learning

127. Handling Rapid Guessing

Paper Session
11:30 to 1:00 pm PT
Westin San Diego Gaslamp: California Ballroom C

Participants:

Comparing the Robustness of Three Nonparametric DIF Procedures to Differential Rapid Guessing

Mohammed Abulela; Joseph A. Rios, University of Minnesota

The robustness of three nonparametric DIF procedures to type I error were examined in the presence of differential subgroup rapid guessing. Sample size, group impact, test difficulty, and differential RG rate were manipulated in the simulation study. To support simulation results, an applied analysis was conducted using PISA science assessment.

Examination of Individual Ability Estimation and Classification Accuracy Under Rapid Guessing Misidentifications

Joseph A. Rios, University of Minnesota

This simulation study investigates the effect of rapid guessing (RG) misclassifications on individual examinee ability estimate bias and proficiency level classification accuracy when using effort-moderated scoring. This is done by manipulating simulee ability level, RG characteristics, and misclassification type and percentage.

Performance Decline as an Indicator of Generalized Test-Taking Disengagement

Steven Wise, NWEA; G. Gage Kingsbury, NWEA

This study explored the stability of student test-taking performance. Using data from a computer-based interim test, 10% of students exhibited meaningfully decreasing performance. Performance decline, which indicated the presence of generalized disengagement, frequently occurred in the absence of rapid guessing and suggests partial test-taking engagement.

To What Degree Does Rapid Guessing Distort Test Performance? A Meta-analytic Investigation

Joseph A. Rios, University of Minnesota; Jiayi Deng, University of Minnesota; Samuel Dale Ihlenfeldt, University of Minnesota

This meta-analysis sought to quantify the average degree of score distortion in test performance due to rapid guessing (RG). Included studies: (a) group-administered a low-stakes cognitive assessment; (b) identified RG via response times; (c) compared filtered and unfiltered data to evaluate the influence of RG on group test performance.

Discussant:

James Soland, University of Virginia

128. Amplifying the Voices of Women of Color in Educational Measurement

Coordinated Paper Session
1:15 to 2:45 pm PT
Westin San Diego Gaslamp: California Ballroom A

This coordinated session features three research studies focused on the experiences and representation of women of color in educational measurement. The studies are designed to amplify the voices of marginalized women in our field, explore the representation of women and people of color on editorial boards of peer-reviewed journals, and examine the importance of mentoring for Black women and Latinas in graduate school. Dr. Susan Lyons will open the session by emphasizing the urgent need to the recruit and retain a more diverse set of measurement professionals. The research will then be presented by three Women in Measurement Fellows, Dr. Jade Caines Lee, Dr. Tomoe Kanaya, and Reka Barton. The session will close with commentary of renowned scholar, Dr. Gerunda Hughes, who will provide her reflections on the significance of the research presented, drawing from her expertise and own experiences in the field.

Session Organizer:

Susan Lyons, Lyons Assessment Consulting

Participants:

Amplifying Marginalized Voices: The Experiences of Black Women in Educational Measurement

Jade Caines Lee, Clark Atlanta University

The Representation of Women and People of Color in the Editorial Boards of Peer-Reviewed Journals for Educational Measurement
Tomoe Kanaya, Claremont McKenna College

Testifying & Testimonios: The Importance of Mentoring Relationships for Black Female and Latina Graduate Students on the Pathway to the Professoriate
Reka Barton

Discussant:
Gerunda Hughes, Howard University

129. Using Sequence-Based Methods on Process Data in Large-Scale Assessments

Coordinated Paper Session
1:15 to 2:45 pm PT
Westin San Diego Gaslamp: California Ballroom B

This symposium highlights four novel studies that draw on sequence-based methods for advancing understanding of process data in international and national large-scale assessments. Process data document the entire process performed by a test-taker to complete a task and therefore hold great potentials in providing new validity evidence and in-depth insights into respondents' cognitive and behavioral processes. However, due to their complex structure, leveraging this potential is not straightforward. In this session, the first paper introduces a machine learning procedure that leverages early-window clickstream data to investigate early predictability of failure on interactive tasks. The second paper presents a sequence mining approach to dynamically track students' keystrokes in mathematical tool usage across multiple tasks. The third paper exhibits an extended hierarchical model that draws on process data in form of eye movement sequences to decompose the duration of sequentially occurring components during the response process and investigate the relationship between speed on the components and latent ability. The fourth paper highlights a dynamic sequential clustering method to investigate students' navigation behaviors in reading tasks. Considering a broad range of process data, these studies show how sequence-based methods may aid in taming this complex data, while taking their sequential character into account.

Session Organizers:
Qiwei He, Educational Testing Service
Esther Ulitzsch, IPM

Chairs:
Qiwei He, Educational Testing Service
Esther Ulitzsch, IPM

Participants:

A Machine Learning Procedure for Early Prediction of Failure
Esther Ulitzsch; Vincent Ulitzsch, Technical University Berlin; Qiwei He, Educational Testing Service; Oliver Luedtke, IPN – Leibniz Institute for Science and Mathematics Education

Using Sequence Mining to Explore Mathematical Tool Usage in Large-Scale Digitally-Based Assessments
Yang Jiang, ETS; Gabrielle Cayton-Hodges, Educational Testing Service; Leslie Nabors Olah, Educational Testing Service; Ilona Minchuk, Educational Testing Service

Decomposing Time-on-Task: An Extension of the Hierarchical Model
Tobias Deribo, DIPF | Leibniz Institute for Research and Information in Education; Ulf Kroehne; Frank Goldhammer, DIPF | Leibniz Institute for Research and Information in Education, ZIB

Clustering reading navigation sequences with dynamic time warping method
Qiwei He, Educational Testing Service; Francesca Borgonovi, OECD; Javier Suarez-Alvarez, OECD

Discussant:
Frank Goldhammer, DIPF | Leibniz Institute for Research and Information in Education, ZIB

130. eBoard Session

Electronic Board Session
1:15 to 2:45 pm PT
Westin San Diego Gaslamp: San Diego Ballroom

Participants:

Methods for exploring test security concerns about extended testing window
Aijun Wang, FSBPT; Yu Zhang, Federation Of State Boards of Physical Therapy; Lorin Mueller, Federation of State Boards of Physical Therapy

Since the breakout of the COVID-19, testing programs have adopted various measures to mitigate its impact. One method is to extend the testing window to allow more candidates to be tested. This paper explores the methods for test security issues related to extending testing window.

Comparability of Paper and Online Testing: A Meta-Analysis of 2010–2020 Research

Ann Arthur, ACT; Jeffrey Steedle, ACT, Inc.; Shalini Kapoor, ACT, INC.

Testing programs transitioning from paper to online testing must conduct research to support claims about score interpretation and use. This paper presents a meta-analysis of mode comparability research for studies published between 2010 and 2020 and compares the findings to previous mode comparability research.

Establishing a Strong Program of Validity for the IXL Real-Time Diagnostic Assessment

Bozhidar M. Bashkov, IXL Learning; Christina Schonberg, IXL Learning; Xiaozhu An

We build upon prior research to establish a strong program of validity for the IXL Real-Time Diagnostic, a formative/interim PreK-12 math and ELA assessment. New data revealed strong correlations of Diagnostic and SOL scores, and high overlap in classification of students into proficiency levels in both math and ELA.

A Comparison of the Testlet Model Against Traditional Approaches

Brian French, Washington State University; Shenghai Dai, Washington State University

Local item dependence (LID) is common, yet a challenge to model. This simulation study examined the accuracy to detect LID with testlet, bifactor, and IRT models, and 3 LID statistics (Q3, □2, and G2)for testlet data. Results support the use of a testlet model to capture LID with accuracy.

Measuring high school curriculum: How do different methods compare?

Burhan Ogut, American Institutes for Research; Ruhan Circi, American Institutes for Research; Darrick Yee, American Institutes for Research

There is no clear consensus on how to measure high school coursework. This study aims to examine how different measures of coursework relate to each other and students' postsecondary enrollment. We compare results from three measures: index-based coursework intensity, cluster analyses, and latent class analysis.

A Cohort Change Model for Longitudinal Analysis of Anonymous Student Surveys

Carlos Chavez, University of Minnesota - Twin Cities; Michael C. Rodriguez, University of Minnesota

The measurement of social emotional learning is an important area of interest in education research. Some limitations in this field of research include methods of assessment and lack of longitudinal data. This study investigates the capacity to model anonymous student responses over time when the data are nested.

Engaging Stakeholders to Develop Narrative Vignettes as Reports of Student Assessment Outcomes

Chad Gotch, Washington State University; Mary Roduta Roberts, University of Alberta; Marcus Poppen, Washington State University; Paul Strand, Washington State University- Tri Cities; Bruce Austin, Washington State University; Brian French, Washington State University

In this presentation we demonstrate robust application of theory, articulations of intended and realized outcomes, and a method for stakeholder engagement to develop of a novel form of reporting results—narrative vignettes. This work complements traditional forms of score reporting to achieve more positive consequences from the testing program.

Comparing Item Features with and without Text Complexity to Model Reading Comprehension

Christina Schneider, Cambium Assessment; Jing Chen, NWEA

This study investigates the comparative influence of text complexity measures to item level metadata on the predictions of the difficulty of the reading comprehension items. Implications for different theories of reading comprehension are discussed.

Exploring Online Item Format Biases Prompted by Common Method Variance: Multidimensional Rasch/Bifactor Approach

Daeryong Seo, Pearson; Insu Paek, Florida State University; Se-Kang Kim, Fordham University

Both general math ability and online item format biases caused by common item variance were estimated using multidimensional Rasch/bifactor approach; no correlation among general ability factor item-type factors was estimated. Both gender and ethnicity variables explained the general ability as well as the biases caused by the online item format.

Detecting Item Parameter Drift in Small Sample Equating

Daniel Jurich, National Board of Medical Examiners; Chunyan Liu, National Board Of Medical Examiners

Screening items for parameter drift helps yield accurate equating outcomes. However, few studies have investigated methods to detect item parameter drift in small sample equating. This study demonstrates that several newly researched drift detection strategies can improve equating accuracy under various conditions with small samples.

The Role of Item Difficulty in Student Engagement Measurement: A Closer Look

Daria Gerasimova, University of Kansas; Angela Miller, George Mason University; Margret Hjalmarson, George Mason University

Using unidimensional engagement subscales and multiple estimation methods, we found that the emergence of the Difficulty Factor in a larger engagement instrument is unlikely to be due to model complexity, non-linearity, or overextraction. Thus, the results provide further evidence that internal structures of engagement constructs are affected by item difficulty.

Impact of Composite Creation in Detecting DIF in Longitudinal Growth Curve Models

Dubravka Svetina Valdivia, Indiana University; Montserrat B Valdivia Medinaceli, Indiana University Bloomington; Shimon Sarraf, Indiana University

Use of composites is common practice in fitting longitudinal latent growth curve models. Composite formation is typically based on either CTT, IRT, or FA. Our simulation study aims to explore the impact of composite creation on longitudinal model's performance in detecting longitudinal measurement noninvariance/DIF.

Latent Group Identification for Automatically Generated Items by Applying Mixture Rasch Model

Eunbee Kim, Georgia Institute of Technology; Susan Embretson, Georgia Institute of Technology

The current paper aims to examine heterogeneity of examinees for math items generated by an automatic item generator. We expect to detect latent groups based on problem-solving strategy. We plan to identify the commonalities of examinees' response characteristics for generated items and compare latent groups for generated items with latent groups of operational items that went through classical item screening procedure.

The Effects of COVID-19 on English Language Development of English Learners

Eunhee Keum, UCLA/CRESST; Nami Shin, UCLA/CRESST; Edynn Sato, UCLA/CRESST; Kilchan Choi, CRESST/UCLA; Yun-Kyung Kim, University of California – Los Angeles

This study examines the impacts of the pandemic and learning disruptions on English learner (EL) students' academic English language development. Using student- and item-level data of the ELPA21 summative assessments, this study compares EL students' performance in two consecutive years and examine if their performance differs between pre- and post-pandemic periods.

Latent class analysis of change for a longitudinal assessment

Fen Fan, NCCPA

The goal of this study is to examine subset of examinees with dis/similar learning trajectory over time and figure out what variables helps explain the different learning trends over time. Findings from the study can help identify features of examinees needing support and help them succeed in the longitudinal program.

Issues with Estimation of Test Information with Underspecified IRT Models

William Skorupski, Data Recognition CORP; Lisa Keller, University of Massachusetts

The purpose of the study is to discuss the importance of model-data fit when estimating the reliability of test data. Several theoretical and empirical scenarios are presented to demonstrate the effect of poor model-data fit on recovery of test information. Results show that underspecified models tend to underestimate true reliability.

Investigating Characteristics of Criterion Scores to Improve Matching for Comparability Studies

Xin Li, ACT, INC.; YoungWoo Cho, ACT

When the implementation of random assignment is impossible, match sampling can be used as an approximation to the randomization. Extend from real data analysis results carried out by the study, a simulation study is also conducted to evaluate what characteristics of a criterion score can improve matching quality.

Bayesian Growth Mixture Models for Classifying and Measuring Individual Trajectories

Xingyao Xiao; Sophia Rabe-Hesketh, University of California, Berkeley

Growth-Curve Modeling (GCM) is a tool for measuring interindividual differences in intraindividual change on a continuum. Growth Mixture Modeling (GMM) adds the capability to examine heterogeneity in the between-individual trajectories due to unobserved classifications of individuals. We introduce GMMs and demonstrate Bayesian estimation in the Stan package.

Optimal item pool characteristics for constrained classification testing with Linear-on-the-fly model

Xuechun Zhou, Pearson

This study examines characteristics of item pools generated under varying conditions for Linear-on-the-fly (LOFT) testing model and their impact on exam performance. All pools are determined through simulations satisfying predetermined constraints using an existing master pool. Simulated and operational pools are evaluated by master pool utilization and classification accuracy.

Item-Focused Trees for DIF Detection in Partial Credit Model

Wei Xu, American Board of Anesthesiology

The current study evaluates the effectiveness of item-focused tree approach when detecting DIF in partial credit model. This approach was compared with alternative approaches in DIF detection in polytomous items. Simulation results show the proposed method is an effective way to detect DIF. This approach might have great implications for psychometric research and practices.

131. Bayesian Methods

Paper Session
1:15 to 2:45 pm PT
Westin San Diego Gaslamp: Del Mar

Participants:

A Bayesian Multivariate Multilevel Modeling Approach for Analyzing Classroom Heterogeneity

Alexander Naumann, DIPF | Leibniz Institute For Research and Information in Education; Dimitra Kolovou, University of Teacher Education St. Gallen (PHSG); Jan Hochweber, University of Teacher Education St. Gallen (PHSG); Anna-Katharina Praetorius, University of Zurich

Common measures of teachers' judgment accuracy have significant limitations. They are prone to sampling and measurement error and do not adequately deal with imbalanced and hierarchical data structures. Thus, we propose a Bayesian multivariate multilevel model tailored to overcome these modeling issues when investigating judgment accuracy in multiple domains simultaneously.

Applying Bayesian Method to Develop Valid and Reliable Instruments

Jihang Chen, Boston College; Zhushan Mandy Li, Boston College; Sebastian Moncaleano, Boston College

Evidence for content and construct validity is necessary for developing a high-quality multi-unidimensional test. We proposed a Bayesian method to integrate experts' ratings and participants' data to establish a unified model of validity evidence. This approach can help us get item-to-domain correlations efficiently and overcome the challenge of small samples.

Bayesian Monte Carlo Simulation Studies: Practice and Implications

Allison Ames Boykin, University of Arkansas; Brian C Leventhal, James Madison University; Nnamdi Ezike, University of Arkansas; Kathryn Nicole Thompson, James Madison University

Monte Carlo simulation studies (MCSS) are prevalent in educational measurement. There is little guidance about specific decisions within the eight steps of MCSS, particularly a lack of transparency and recommendations for Bayesian MCSS. This study reviewed seven educational measurement journals, totaling 1230 journal articles, and provides recommendations for Bayesian MCSS.

Compromised Item Detection: A Bayesian Change-point Perspective

Yang Du; Susu Zhang, University of Illinois at Urbana-Champaign; Hua-Hua Chang, Purdue University

Item compromise has been a longstanding issue that threatens test security. We proposed a Bayesian change-point method to detect compromised items. We conducted two simulations and our results showed the post-change model and item change-points can be identified and the detection procedure showed high power and short detection lags.

132. Using GitHub for Open-Source Analytics, Reporting, and Dissemination of Research

Demonstration Session
1:15 to 2:45 pm PT
Westin San Diego Gaslamp: La Jolla

Open-source software has become increasingly popular over the past decade. By making their work open source, researchers can easily develop and distribute tools for data analysis. Importantly, employing open-source software provides numerous advantages for research in education, including (a) being freely available, (b) fostering collaboration among analysts and policymakers, (c) quickly identifying and remedying errors, and (d) promoting reproducible research practices.

Session Organizer:
Damian Betebenner, Center for Assessment

Participants:

Using GitHub for Developing an R Package

Damian Betebenner, Center for Assessment

In the first demonstration, we present a repository skeleton for building and distributing an R package. Perhaps the greatest strength of R is its extensibility, with the ability to expand base R in innumerable ways. In this demonstration, we present a way to quickly build your first R package from a package skeleton on GitHub.

Web-based Distribution of Research Using GitHub

Allie Cooperman, University of Minnesota

The second demonstration shows how researchers can collate their analyses into a user-friendly, distribution-ready website. Combining GitHub's repository support for a static website (GitHub Pages) and author-written R packages (also available through GitHub), researchers can host vignettes, source code for statistical analyses, presentations, and project updates.

Using GitHub to Automate Report Generation

Adam VanIwaarden, Center for Assessment

Data analysts are often tasked with writing reports that describe data, analyses, and results associated with a project. Depending upon the nature of the project, such reports are either completely customized or borrow heavily from other reports GitHub repositories can be used to coordinate the writing and the production of final reports for dissemination.

Doing it All in a Single GitHub Repository

Nathan Dadey, Center for Assessment

The last demonstration illustrate how researchers can use a single GitHub repository to design, analyze, and disseminate educational research from start to finish. This omnibus repository incorporates a suite of tools, such as author-written R packages and automated report generation, that can be easily customized to fit a particular project.

133. Issues in Remote Testing During the Pandemic

Coordinated Paper Session
1:15 to 2:45 pm PT
Westin San Diego Gaslamp: Santa Fe

The COVID-19 pandemic forced many students to learn, and take assessments, away from the school building, oftentimes in an uncontrolled environment. In this session we will review several issues encountered in remote testing during the 2020-2021 school year. The first paper discusses who was testing remotely and why. This paper examines demographic differences between remote and in-school testers and uses a four-level hierarchical linear model to examine the associations between testing in-school and several student-, school-, district-, and state-level predictors. The next two papers focus on data quality, with the second paper reviewing attempts to parse out good and bad remote testing data, while the third paper advocates for a longitudinal approach to examining data quality, as opposed to traditional cross-sectional methods. Finally, we conclude with a look at Fall 2021 achievement differences based on where students tested during the 2020-2021 school year. Specifically, we examine how post-pandemic performance differed for students who spent the entire 2020-2021 school year testing remotely, compared to those who were in-school for all or part of the school year.

Session Organizer:
***Logan Rome*,** Curriculum Associates

Participants:
Factors Related to Testing Location During the 2020-2021 School Year
Logan Rome, Curriculum Associates; Luciana Cancado, Curriculum Associates

School Versus Remote Testers: Rating the Consistency of Test Scores
Kristin M. Morrison, Curriculum Associates; Kevin Cappaert, Curriculum Associates; Laurie Davis, Curriculum Associates

Disrupted Data: Using Longitudinal Assessment Systems to Monitor Test Score Quality
Lily An, Harvard Graduate School of Education; Andrew Ho, Harvard Graduate School of Education; Laurie Davis, Curriculum Associates

Student Growth During COVID
Matt Dawson, Curriculum Associates

Presenters:
Kristin M. Morrison, Curriculum Associates
Lily An, Harvard Graduate School of Education
Matt Matt Dawson, Curriculum Associates

Discussant:
Derek Briggs, University Of Colorado

134. Improving Fairness Evaluations in Automated Scoring

Coordinated Paper Session
1:15 to 2:45 pm PT
Westin San Diego Gaslamp: Plaza

The Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014) advocate for reviewing automated scoring algorithms for potential sources of bias. At present, however, there is a lack of industry consensus as to how to evaluate automated scoring at the subgroup level. Previous efforts have often been limited by unstable bias estimates, emphasis on item-level rather than engine-level evaluations, the need to control for ability in order to examine comparable groups, and engine complexity (Gregg, Young, & Lottridge, 2021). In this symposium, two vendors will describe novel approaches to examining automated scoring for bias developed in response to these limitations. Specifically, four different approaches will be described, using such diverse techniques as propensity score matching, conditioning on the score point distribution, linear modeling, and multilevel multinomial logistic regression. Presenters will share results of these approaches applied to interim and summative assessment data. The session will conclude with a discussion from a psychometric and state perspective on fairness evaluation in the context of automated scoring. Participants can expect to leave the session with an understanding of alternate approaches to bias evaluation designed to better support valid score interpretations.

Session Organizer:
Corey Palermo, Measurement Incorporated

Participants:
A Propensity Scoring Matching Approach
Susan Lottridge, Cambium Assessment, Inc

Conditioning on Score Point Distribution Approach
Arianto Wibowo, Measurement Incorporated

A Linear Model Approach
Derek Justice, Measurement Incorporated

A Multilevel Multinomial Logistic Regression Approach
Yong He, Measurement Incorporated; Shumin Jing, Measurement Incorporated; Yang Lu, Measurement Incorporated

Discussant:
***Liru Zhang*,** Assessment Consulting Services

135. Response Time Models

Paper Session
1:15 to 2:45 pm PT
Westin San Diego Gaslamp: California Ballroom C

Participants:

Understanding response process heterogeneity across fast and slow responses in noncognitive tests

Nana Kim, University of Wisconsin, Madison; Daniel Bolt, University Of Wisconsin, Madison

This study investigates the heterogeneity in response processes within individuals in noncognitive tests in relation to response times. We examine the differences in item characteristics across fast and slow responses and investigate how response times associate with heterogeneous response processes involving response styles.

Detecting Careless Responses with the Rapid Guessing-Slow Responding Method

Okan Bulut, University of Alberta; Seyma N. Yildirim-Erbasli, University of Alberta

We propose a new method for handling careless item responses by taking response accuracy into account. The proposed method is compared with the existing method in a simulation study. Results suggest that the proposed method is a better solution to reduce the impact of data contamination because of careless responding.

Identifying Examinee-Level Speededness on Cognitive Tests Using Factor Score Estimate

Tanesia Beverly, Google, LLC; Alexander Weissman, Law School Admission Council; Eric Eric Loken, University of Connecticut

Summary statistics used to evaluate test speededness are not designed to identify speeded examinees. They provide the degree of speededness on timed tests across examinees. Therefore, a novel approach for identifying examinee-level speededness is introduced and compared with mixture modeling. Results suggest both methods perform similarly and are effective.

Item Fit Indices for Continuous Response Time Models

Weicong Lyu, University of Wisconsin - Madison; Xiang Liu, Educational Testing Service; Sandip Sinharay, Educational Testing Service; Matthew Johnson, ETS

We introduce a class of fit indices based on a conditional test approach for assessing item level fit of response time models. The type-I error rate and statistical power are examined through simulation studies. In addition, the utility of the method is demonstrated by analyzing a real dataset.

Modeling the Sequential Response Time with Item Position and Total Time Limits

Yi Chen, Teachers College Columbia University; Sizheng Zhu, Teachers College Columbia University; Yi Yang, Teachers College Columbia University; Young-Sun Lee, Teachers College Columbia University

In this study, we proposed the general framework of modeling sequential RTs. This framework extending the conventional treatment of RTs by including the truncation of response distribution and adding the item position effects. These two treatments give a more realistic description of RT generation. Both simulation study and real data analysis are conducted via a fully Bayesian approach with Markov Chain Monte Carlo (MCMC) method.

Discussant:

Chun Wang, University of Washington

We’re Here Too: Researchers from Historically Marginalized Groups Networking Event

3:00 to 4:30 pm PT
Westin San Diego Gaslamp: Garden Terrace

Collaborative social event hosted by CODIT, GSIC, Membership Committee and Outreach Committee

136. Business Meeting and Presidential Address

4:40 to 6:30 pm PT
Westin San Diego Gaslamp: California Ballroom ABC

137. Presidential Reception (Open to all)

6:30 to 9:00 pm PT
Westin San Diego Gaslamp: Garden Terrace

138. NCME Fitness Run/Walk

6:00 to 7:00 am PT
Meet in lobby of Westin San Diego Gaslamp

139. Opportunity-to-learn as a means to enhance equity and increase understanding

Coordinated Paper Session
8:00 to 9:30 am PT
Westin San Diego Gaslamp: California Ballroom A

Opportunity-to-learn has evolved from a focus on whether students have had sufficient access to instruction to a more robust conception regarding the conditions and resources provided to schools to enable students to succeed. The Standards in Educational and Psychological Testing (AERA, APA, & NCME, 2014) conceptualized OTL as 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—has several implications for the fair and valid interpretation of test scores for their intended uses" (p. 56). In response to the pandemic-related schooling disruptions, many state and district leaders felt the sense of urgency to document the impact of COVID-19 and turned to OTL measures to help contextualize test score interpretations for the 2019-2020 school year and to gain a broader understanding of the pandemic effects. We argue that OTL measures should be used every year and we present a conceptual framework for broadly documenting OTL, including indicators of academic learning, social-emotional wellness, and other resources. Additionally, the presenters illustrate—using OTL 2020-2021 data from multiple states—what can be learned from including OTL indicators in regular assessment and accountability reporting.

Session Organizer:
Scott Marion, National Center for the Improvement of Educational Assessment

Participants:

A Justice-orientated Framework for Opportunity-to-Learn Indicator Use
Thao Vo, Washington State University; Scott Marion, National Center for the Improvement of Educational Assessment

Coherence in an Opportunity-to-Learn Indicator System
Daniel Silver, University of Southern California Rossier School of Education

Using Process Data to Understand Fine-Grained OTL
Tanner Jackson, Educational Testing Service

Developing Indicators of Social and Emotional Learning Opportunities
Laura Hamilton, ETS

140. Recruiting and Retaining New Educational Measurement Faculty

Organized Discussion
8:00 to 9:30 am PT
Westin San Diego Gaslamp: California Ballroom B

Several researchers have warned that a shortage of educational measurement graduate students, with sufficient training to fill measurement positions, has the potential to impede the successful development and use of educational assessments within the United States (Randall et al., 2020). This session, hosted by the Educators of Measurement Special Interest Group in Measurement in Education, proposes an engaging, organized discussion around the topic of recruitment and retention of educational measurement faculty. We intentionally emphasize women and those who identify in a minoritized group in the discussion. We will focus on topics such as recruitment into academia, faculty diversity, mentorship of new faculty, and changes to the field that can mitigate the shortage of faculty. The session begins with a video of graduate students and new faculty responding to question prompts around recruitment into academia and retention of new faculty. The small groups will follow a 3-stage brainstorm and discussion format. Audience members will participate in a lively discussion in small groups before returning as one group to share their ideas. Understanding the impediments to recruiting and retaining new faculty, especially minority faculty, will improve the experience of faculty and the graduate students in their programs.

Session Organizer:
Allison Ames Boykin, University of Arkansas

Presenters:
Joseph A. Rios, University of Minnesota
Yi Zheng, Arizona State University
Anne Corinne Huggins-Manley, University of Florida

141. Reimagining Assessments: The Responsibilities is Ours!

Organized Discussion
8:00 to 9:30 am PT
Westin San Diego Gaslamp: California Ballroom C

The pandemic and social justice issues jarred the face of education and assessment. Initial chaos was followed by profound support and leadership. Still, most of our students were negatively impacted socially, emotionally, and academically due to the disruption of our educational system. Most heavily impacted were our Latino, Black, and marginalized students. Digital access and inequities in our current system caused a great deal of concern, raising questions about how we would assess students effectively and more importantly, about the effectiveness of our current assessments altogether. During this session, the NCME State and Local Assessment Leaders (SALAL) will engage in a candid conversation about reimagining assessment. Panelists propose to redefine the purpose for assessment with a common vision. They will call for a comprehensive assessment system based on input from all major stakeholders. Moving beyond administering assessments and compliance, panelists will drive the discussion towards increasing the value of assessments. They are dedicated to understanding and addressing our greatest assessment challenges. Short- and long-term collaborations will focus on partnerships, capacity building, and supporting good local practices of comprehensive and holistic assessment systems that advance learning for all students. With a commitment and plan we can genuinely reimagine our assessments.

Session Organizer:
Elda Garcia, National Association of Testing Professionals

Presenters:
Lisa Sireno, Missouri Department of Elementary and Secondary Education
Stephen Sireci, University of Massachusetts Amherst
Mary Anne Arcilla, Educational Testing Service
Keri Rodrigues, National Parents Union
Maria Armstrong, Association of Latino Administrators and Superintendents

142. Recent Challenges to Ensuring Score Comparability

Coordinated Paper Session
8:00 to 9:30 am PT
Westin San Diego Gaslamp: Del Mar

Scores are the most visible and widely used products of an assessment system. Score users presume that the score means what the score provider claims it means. Although a central claim is that scores obtained on different assessments are equated and can be used interchangeably, equating has been distinguished from other popular forms of score linking. Over the past 15 years, the testing field has experienced considerable changes, such as to testing populations, administrations (i.e., digital), tests and item bank security threats, demand for innovative testing, and to the tests themselves. These recent changes go beyond those addressed in previous texts and raise new questions about score linking, and claims made about test scores and communicated to test users. This proposed coordinated session addresses testing practices and the challenges these present to equating within a testing program, to the linking of tests across testing programs, and to the use and interpretation of test scores. Test linking experts will present five papers addressing specific challenges that that testing field changes present to test linking results. Another expert from the measurement field will discuss and synthesize the perspectives from the papers.

Session Organizers:
Tim Moses, College Board
Gautam Puhan, ETS

Participants:

The Impact of Test Security on Score Linking and Score Comparability
Jinghua Liu, Pearson; Kirk Becker, Pearson

Threats to score comparability with at home testing and how to address them
Gautam Puhan, ETS; Sooyeon Kim, ETS

Historical Perspectives on Score Comparability Issues Raised by Innovations in Testing
Peter Baldwin, National Board of Medical Examiners; Brian Clauser, National Board of Medical Examiners

Linking by Definition and Presumed Invariance Across Conditions of Measurement
Tim Moses, College Board

Score Comparability between in-person and remote proctored exams
Paul Edward Jones, Pearson VUE; Ye Tong, NBME; Jinghua Liu, Pearson; Joshua Borglum, Pearson; Vince Primoli, Pearson

Discussant:
Robert Brennan, University of Iowa

143. Validity Issues

Paper Session
8:00 to 9:30 am PT
Westin San Diego Gaslamp: La Jolla

Participants:

Principles of leadership with validity: A theory of ownership

Ian Hembry, Metametrics

Validity is of central import to any measurement endeavor. The Standards (2014) state that "validity is…the most fundamental consideration in developing tests and evaluating tests." The following argument is an interdisciplinary approach that extends validity theory to incorporate principles of leadership as applied across multiple disciplines.

Situating the validity argument: lessons from feminist epistemologies.

Sergio Araneda, University of Massachusetts Amherst

The argument-based approach to validity (e.g., Kane 2013) is based on a constructionist epistemology. In this presentation I will introduce the concept of situated knowledges (Haraway, 1988) and explain how those ideas can enhance the argument-based approach. Recommendations for acknowledging the situated nature of a validity argument are provided.

Visualizing Validity Evidence: Considering Strength of Evidence Following Disrupted Administration

Amy Clark, ATLAS: University of Kansas; William Jacob Thompson, University of Kansas; Jennifer Kobrin, ATLAS: University of Kansas

Validity arguments consider the strength of evidence for supporting intended interpretations and uses. When instruction and assessment administration are disrupted, test developers must consider the strength of validity evidence and whether intended uses are supported. We demonstrate a validity visualization method for evaluating relative strength of validity evidence.

Discussant:

Martha McCall, McKinsey & Company

144. Focus on Diagnostic Classification Models

Paper Session
8:00 to 9:30 am PT
Westin San Diego Gaslamp: Santa Fe

Participants:

A Multilevel Polytomous-attribute Diagnostic Classification Model

Yu Bao, James Madison University; Nicolas Mireles, James Madison University; Jin Liu, University of South Carolina

Cluster sampling often occurs during test administration. Current literature introduced multilevel DCMs to account for the violation of the sample independent assumption. Further delineating attributes into more than 2 levels may be useful in various scenarios. We propose a multilevel polytomous-attribute DCM to provide finer granularity of the diagnostic feedback.

Application of Diagnostic Classification Models to Diagnosing Misconceptions with Constructed-Response Items

Wenjing Guo, University of Alabama; Louis Roussos, Cognia; William Stout, University of Illinois; Robert Henson, University of North Carolina; Xi Wang, Cognia; Liuhan Cai, Cognia

Diagnostic classification models have been developed for diagnosing misconceptions with multiple-choice items. Presented here is an application of one such model to constructed-response items on an algebra readiness test, identifying new challenges and developing procedures to address them. Initial results indicate a promising new area of research for diagnosing misconceptions.

Comparing DCM and IRT Reliabilities with New DCM Measures and IRT Analogs

Qi Huang, University of Wisconsin - Madison; Daniel Bolt, University of Wisconsin, Madison

To compare the reliability of DCM and IRT scoring, we propose an alternative way of examining IRT-based reliability based on IRT analogs for three new DCM reliability measures consistent with the intent of binary mastery classification. The results suggest that the IRT-based reliabilities generally appear to be higher.

Diagnostic Concept Inventory for Misconceptions in Probabilistic Reasoning

Madeline Schellman; Laine Bradshaw, University of Georgia; Hollylynne Lee, North Carolina State University; Hamid Sanei, North Carolina State University; Jessica Masters; Lisa Famularo, Research Matters, LLC

We present a concept inventory developed using a principled-design approach and diagnostic classification modeling framework. We overview the development process of gathering extensive response process validity evidence and detail empirical evidence supporting reliable misconception classifications and well-performing items. We highlight the advantages of leveraging diagnostic psychometrics for improved instructional relevance.

Discussant:

Susu Zhang, University of Illinois at Urbana-Champaign

145. Linking Product and Process Evidence in Student Writing

Coordinated Paper Session
8:00 to 9:30 am PT
Westin San Diego Gaslamp: Plaza

This session presents an automated trait model for writing, using data from a large (1.37 million submission) corpus of student essays. It exploits NLP features drawn from the e-rater® automated scoring engine, the TextEvaluator® readability engine, and other ETS text analysis tools (Attali & Burstein, 2006; Sheehan, Kostin, Napolitano & Flor, 2014; Burstein, et el., 2018). We identify a factor structure related to text structure (organization, cohesion, elaboration), control of language (sentence length, sentence complexity, vocabulary length, vocabulary difficulty, adherence to conventions), and genre or register (academic language, concreteness, stance-taking [argument style], and contextualization [narrativity]). We combine this model with keystroke logging data that characterize student writing processes, including overall fluency, typing speed, pausing, and editing behaviors. We present several specific studies: • Using the trait model to measure differential patterns of growth before and after instruction • Using the trait model to characterize changes in the writing of students who successfully passed a high-stakes writing examination after an initial failure • Using the trait model to create group profiles, combining both product and process features • Using the trait model to examine relations between student plans created via planning tools and the quality of the corresponding essays.

Session Organizer:
Paul Deane, ETS

Participants:

Overview of the 12-Trait Model
Paul Deane, ETS; Duanli Yan, ETS

Profiling School Differences in Writing Performance Before and After Instruction
Paul Deane, ETS

Profiling Changes in Student Writing: From Failure to Success
Klint Kanopka

Using Trait Profiles to Track Individual and Group Differences on a High-Stakes Writing Task.
Mo Zhang, Educational Testing Service

Profiling Multi-Stage Writing Processes: Linking Plan Quality to Essay Performance
Yi Song, Educational Testing Service

Discussant:
Russell G Almond, Florida State Univeristy

146. Avoiding a Train Wreck: Working with Constituents to Rethink Equity in Assessment?

Organized Discussion
9:45 to 11:15 am PT
Westin San Diego Gaslamp: California Ballroom A

The growing recognition of the need to address longstanding issues of social justice is causing some state education departments to revise content standards and some educational assessment organizations and many members of our field to rethink traditional definitions of equity and fairness. Evolving definitions include in them the need to account for differences in sociocultural backgrounds, funds of knowledge, interests, values, and practices that individuals from diverse cultures bring to learning and assessment. At the same time, political resistance to addressing social-justice issues is growing. States have passed laws prohibiting teaching about systemic racism, conscious and unconscious bias, privilege, discrimination, and oppression. A significant cadre of politicians is working to make critical race theory a midterm, as well as a presidential, campaign issue. It is clear that our rethinking the meaning and implementation of equity in assessment will be welcomed by some and rejected by others. Given this deeply divided context, how do we work with our constituents—state policy makers, educators, parents, and the public—in constructive ways to effect a rethinking? How do we avoid a backlash—the train wreck--that causes the field to retrench?

Session Organizer:
Randy Bennett, ETS

Presenters:
***Ye Tong*, NBME**
Michael C. Rodriguez, University of Minnesota
Emi Iwatani, Digital Promise
Derek Briggs, University of Colorado

147. Investigating and Addressing Bias in Board Certification Exams

Organized Discussion
9:45 to 11:15 am PT
Westin San Diego Gaslamp: California Ballroom B

It is easy to agree that certification examinations should be free of bias leading to differences in performance between race/ethnicity groups. What can be more complicated is operationalizing a process that addresses: eliminating problematic content during the test development phase; detecting problematic content using statistical procedures post-administration; accounting for problematic content when making scoring decisions; and using the information from statistical analyses to revise questions to prevent the same bias being demonstrated in the future. In this session, staff from three national certifying bodies will discuss their processes for investigating and addressing bias in their exams. The session will touch on each organization's successes, challenges, and lessons learned. The session will then conclude with a discussion about how addressing bias in exams fits into the bigger picture of diversity, equity, and inclusion (DEI), including understanding the limitations of the current processes described in this session, and brainstorming on what future work may be done to increase the chances of a positive impact. Post-session, attendees should better understand the different processes an organization can implement to reduce bias in their exams and the vital (and sometimes difficult) questions which must be asked when addressing this important topic.

Session Organizer:
Ying Du, American Board of Pediatrics

Moderator:
Ting Wang, American Board of Family Medicine

Presenters:
Karen Hoeve
Kevin Joldersma, American Board Of Emergency Medicine
Mary M. Johnston, American Board Of Emergency Medicine
Thomas O'Neill, ABFM

148. Psychometric Considerations in the Measurement of Social-emotional Learning and School Climate

Coordinated Paper Session
9:45 to 11:15 am PT
Westin San Diego Gaslamp: California Ballroom C

While measures of social-emotional learning (SEL) and school climate are growing in prominence as components of schools' accountability and improvement systems, much less is known about the psychometric quality of the instruments used in practice. In particular, understanding the development of students' perceptions of SEL and climate requires modeling and interpreting growth trajectories, yet little is known about how much common problems with Likert-based measures affect estimates of growth. In this panel, we examine the intersection between self-report issues and a desire to estimate developmental trajectories in two ways. First, we examine the assumptions required to scale SEL and school climate measures for growth inferences, including the appropriateness of various item response theory models to produce scale scores for use in growth models, longitudinal measurement invariance assumptions, and whether the stability of scores is due in part to consistent response styles. Second, we consider the promise and limitations of alternatives to self-report Likert-type items for measuring SEL, in particular using open-ended item responses and text-mining methodologies. Altogether, this session can be useful to measurement experts, practitioners, and policymakers alike by illuminating how much faulty assumptions about self-report measures might affect growth estimates and briefly considering design- and psychometric-based alternatives.

Session Organizer:
Megan Kuhfeld, NWEA

Participants:

Measuring Growth in Students' Social-emotional Learning: A Comparison of Multiple Scoring Approaches
Megan Kuhfeld, NWEA; James Soland, University of Virginia

Multilevel Longitudinal Measurement Invariance and School Climate Surveys
Jon Schweig, RAND

Accounting for Students' Socially Desirable Responding in the Measurement of Social-emotional Skills
James Soland, University of Virginia; Megan Kuhfeld, NWEA

Using Text-Mining Models to Quantify Creative Thinking
Denis Dumas, University of Denver

Discussant:
Daniel Bolt, University Of Wisconsin, Madison

149. Research Blitz

Research Blitz Session
9:45 to 11:15 am PT
Westin San Diego Gaslamp: Del Mar

Chair:
Anthony Albano, University of California, Davis

Participants:

Detecting Speededness Among Examinees Using Change Point Analysis

Sanaz Nazari; J. B. Weir, National Commission on Certification of Physician Assistants; Joshua Goodman, NCCPA

Time limits in certifying exams may introduce another factor effective on the item and ability estimates which has not been accounted for and lead to biased parameters. Change point analysis (CPA) can be used to obtain more accurate parameters by detecting speeded examinees, change point, and removing speeded responses.

Differences in time usage as a competing hypothesis for gender differences in PISA 2018

Radhika Kapoor; Erin Michelle Fahle, St John's University; Ana Carolina Trindade Ribeiro, Graduate student; David Jose Klinowski Gomez, Graduate student; Ben Domingue, Stanford University

We posit and analyze an alternative candidate explanation for gender differences in performance on PISA 2018. We study the degree to which these differences may be due to differences in time usage behaviors, a viable alternative given the exam's low-stakes and gender differentials in effort and risk aversion.

Evaluating Distractor Quality Through Eye Tracking

Victoria Yaneva; Brian Clauser, National Board of Medical Examiners; Janet Mee, NBME; Amy Morales, National Board of Medical Examiners; Miguel Paniagua, National Board of Medical Examiners

We investigated whether eye-tracking data can be used to detect ineffective distractors to allow for their replacement or removal prior to pretesting. The results suggest that these data can be used to identify distractors that would rarely—if ever—be selected by test takers responding in a high-stakes setting.

Improving the Practicality of Automatic Essay Scoring Systems Using Active Learning Models

Tahereh Firoozi, University of Alberta; Mark Gierl, University of Alberta

This study investigated the performance of machine learning automatic essay scoring (AES) systems with less training data in two different languages using active learning models. The results indicated that the AES systems can perform equally accurately even with less training data with high variation. The results will inform the use of AES systems in classroom context and online learning courses.

Evaluating Person-fit Based on Joint Likelihood of Responses and Response Times

Suhwa Han, University of Texas at Austin

The study evaluates a person-fit statistic based on the joint likelihood of item responses and response times. Using simulated data as well as empirical data, the study in particular investigates the robustness of the person-fit results when the empirical response time distribution does not follow the presumed model.

Standard Setting: examining the accuracy of proficiency scores in creating cut-scores using Receiving Operating Curve (ROC)

Dongwei Wang, UMass Amherst; Lisa Keller, University of Massachusetts

The standard in most licensure and certification examinations are currently using modified Angoff method. In medical related fields, receiving operating characteristic (ROC) is commonly used to create optimal cut-scores to minimize the error rate. This study compared the error rate between theta score scale using IRT and raw score scale.

The Impact of Test Speededness on Scale Stability

Hongling Wang, ACT, INC.; Dongmei Li, ACT

This study investigated the impact of test speededness on scale stability for equating using the random groups design. The impact of changes in a few other factors, including test population ability level, test form difficulty consistency, and equating methods, were also investigated. Preliminary results showed little impact of test speededness.

150. Score Resolution Rules in Constructed Response Scoring

Coordinated Paper Session
9:45 to 11:15 am PT
Westin San Diego Gaslamp: La Jolla

The scoring of constructed responses often involves more than one rating to ensure that the impact of rater errors is minimized and reported scores reflect the skills of the test taker. In some cases, a subset of responses is selected to be double scored and in other cases all responses received multiple ratings. Some recent research suggests that these scoring decisions should be carefully considered and evaluated as they may yield different levels of measurement error (e.g., Cohen, 2015). In this session, four papers discuss score resolution methods under different contexts. The first paper applies statistical decision theory to determine which test takers should be targeted for double scoring on a performance assessment. The second paper uses a simulation study to evaluate adjudication practices in the double-human scoring context with varying degrees of bias and variability in the raters and the adjudicator. The third paper also focuses on adjudication in the double-human scoring context, but relies on empirical data to estimate rater bias and identify cases in which the rater unnecessarily triggered an adjudication. Finally, the last paper uses data from a large-scale assessment to compare score resolution methods in the context of human and machine score combinations.

Session Organizer:
Jodi Casabianca-Marshall, Educational Testing Service

Participants:
Targeted Double Scoring of Performance Tasks Using a Decision-Theoretic Approach Sandip
Sinharay, Educational Testing Service; Wei Wang, ETS; Matthew Johnson, ETS; Jing Miao, Educational Testing Service

Evaluation of Adjudication Rules for Essay Scoring: A Monte Carlo Simulation
Nnamdi Ezike, University of Arkansas; Michael E. Walker, Educational Testing Service; Jodi Casabianca-Marshall, Educational Testing Service

Adjudication: Who Needs It?
Michael E. Walker, Educational Testing Service; Sooyeon Kim, ETS

Optimally Reconciling Human and Machine Score Discrepancies
Jodi Casabianca-Marshall, Educational Testing Service; Wei Wang, ETS

Discussant:
Yoav Cohen, נוף הרים 106

151. English Language Learners

Paper Session
9:45 to 11:15 am PT
Westin San Diego Gaslamp: Santa Fe

Participants:
A Latent Profile Analysis of Academic and Language Proficiency of English Learners
Hanwook Yoo, Educational Testing Service; Mikyung Kim Wolf, Educational Testing Service; Laura D Ballard, Educational Testing Service
Using data from two states, we conducted latent profile analysis to investigate the relationship between English learner (EL) students' performance on a high-stakes content assessment and two English Language Proficiency (ELP) assessments. We compared students' latent profiles estimated by content-assessment performance (predicted) to their classifications by the ELP assessments (observed).

Evaluating Correspondence between English Language Proficiency (ELP) Standards and Academic Content Standards
Kelley Wheeler, ACS Ventures, LLC; Susan Davis-Becker, ACS Ventures, LLC
States are required to implement English Language Proficiency standards that correspond to their content standards; however, there is limited guidance for conducting this process. This presentation will detail the process and evaluate the results of a correspondence study that was designed based on guidance provided by the CCSSO (2012) Framework.

Exploring Greater Efficiencies in Testing: ELs' Language Proficiency and Summative Standardized Assessments
Yen Vo, University of Iowa; Heather Rickels, The University of Iowa; Catherine Welch, University of Iowa; Stephen B. Dunbar, University of Iowa; Annette Vernon, University of Iowa
This study examined the relationship between EL performance on an EL proficiency assessment and their achievement on a statewide summative assessment. The results showed strong classification consistency across the two assessments although their internal structures were somewhat different.

Understanding the Impact of Disrupted Schooling on English Learner Student Achievement
Nami Shin, UCLA/CRESST; Jenny Kao, UCLA CRESST; Edynn Sato, Sato Education Consulting LLC
This paper presents a framework for understanding and evaluating the impact of disrupted schooling on English learner (EL) student achievement. It describes factors that have particular bearing on EL students and that local and state education agencies can directly address through instruction, services, and/or programs.

Discussant:
***Mark Hansen*, **UCLA

152. Advances in Dichotomous IRT

Paper Session
9:45 to 11:15 am PT
Westin San Diego Gaslamp: Plaza

Participants:
A Strategy for Estimating IRT Proficiency Standard Errors Directly
Peter Baldwin, National Board of Medical Examiners
A standard error is the standard deviation of an estimator's sampling distribution and in this paper, it's shown how to calculate the standard deviation of a proficiency estimator's sampling distribution directly for the 1-parameter logistic IRT model. A simulation study found that the proposed method outperformed Fisher-information-based standard errors.

Evaluating Item Parameter Recovery with the Sequential 2PL-IRT Model in Unstructured Data

Ziying Li, University of Florida; Anne Corinne Huggins-Manley, University of Florida; Walter Leite, University of Florida

For advancing personalized learning algorithms in virtual-learning-environments, it is critical to have unbiased item parameters underlying the assessment items. In this study, we evaluate item parameter recovery of the sequential 2PL-IRT model in unstructured data which contains different proportions of multiple-attempt item responses, variability of ability growth, and missingness.

Sequential Item Response Model for Multiple-Choice, Multiple-Attempt Test Items

Yikai Lu, University of Notre Dame; Ying Cheng, University of Notre Dame

A 3PL based sequential item response theory (SIRT) model for the answer-until-correct procedure is proposed. The new model can take guessing into account at each trial and improve person parameter estimates compared to existing models, including the 3PL model and the Rasch SIRT models.

What Happens when Heterogeneity is Added to the 2PL IRT Difficulty Parameter?

Alexandra Lane Stone, University of Connecticut; Eric Eric Loken, University of Connecticut

Following Kelderman and Molenaar's factor analysis demonstration, we examine person specific measurement models in IRT. We simulate person specific item and test level random effects for the item difficulties, and show they are largely invisible in typical model diagnostics, despite the major violation of standard model assumptions.

Fitting psychometric framework to digital-first assessment

Selena Wang; J.R. Lockwood, Duolingo; Yigal Attali, Duolingo

Generation of Duolingo English Test (DET) items are based on AI algorithms. We propose two IRT approaches to model such data and compare their performances using simulations. Our study contributes to psychometric research taking into account biased selection of samples as well as the modeling of digital test items.

Discussant:

Peter Halpin, UNC-Chapel Hill

153. Approaches to Missing Data

Paper Session

1:15 to 2:15 pm PT

Westin San Diego Gaslamp: California Ballroom A

Participants:

Modifying the M2 Statistic to Handle Missing Data

Jeffrey Hoover, University of Kansas; William Jacob Thompson, University of Kansas

We modified the M2 statistic to relax the requirement of no missing data. This study compares the Type I error and statistical power of the modified M2 to the unmodified M2 where the requirement of no missing data is met by coding missing data as incorrect responses and listwise deletion.

A Comparison of Parameter Estimation Algorithms for MIRT Models with Missing Data

Jiaying Xiao, University of Washington; Chun Wang, University of Washington

This study proposed the Gaussian variational EM algorithm with bootstrap bias correction (GVEM-BS) and compared its estimation precision to the Metropolis-Hastings Robbins-Monro (MH-RM) algorithm in multidimensional item response theory (MIRT) models. The simulation results demonstrated the robustness and estimation precision of GVEM-BS in the context of high missing proportions.

Comparing Imputation Methods for School-Level Variables in Large-Scale Assessments

Xiaying Zheng, American Institutes For Research; Sinan Yavuz, American Institutes for Research; Yifan Bai, American Institutes for Research; Markus Broer, American Institutes For Research

When missing data occurs in the school-level variables in large-scale assessments such as NAEP, it creates considerable complications for imputation. This study compares two multiple imputation approaches: chained equations using R and Blimp packages via a simulation study. The results provide empirical suggestions for researchers dealing with multilevel missing data.

Discussant:

Chunling Niu, University of Kentucky

154. Standard Setting

Paper Session

1:15 to 2:15 pm PT

Westin San Diego Gaslamp: California Ballroom B

Participants:

A Standard Setting Method for Cluster-Based Assessments: The AMP Method

Widad Abdalla; Frank Rijmen, Cambium Assessment, Inc

Standard setting "refers to the process of establishing cut scores on examinations" (Cizek, 2006, p. 225). The purpose of this study is to describe a new standard setting method for cluster-based assessments along with a statistic that measures rater stability, and present results of a simulation study.

Bad Standard Setting – and What to Do about It

Gregory Cizek, University Of North Carolina

Standard setting is the most visible and consequential activity for any testing program. However, contemporary standard setting can occur in challenging situations that necessitate deviations from best practices. This presentation addresses five common challenges, their origins, threats posed to the validity of results, and proposes effective strategies for addressing them.

Predictive Utility of a Proficiency Cut Score in a Benchmark Assessment

Takeshi Terada, Arizona State University

This study assessed whether a cut score in a benchmark assessment maximizes the accuracy and minimizes the inaccuracy in predicting the proficiency level in the state assessment. By applying a precision-recall curve (PRC), this study discusses potential approaches to increase the predictive accuracy of the cut score in benchmark assessments.

Discussant:

Scott Marion, National Center for the Improvement of Educational Assessment

155. Focus on Multistage Testing

Paper Session

1:15 to 2:15 pm PT

Westin San Diego Gaslamp: California Ballroom C

Participants:

An Approach to Assembling Multistage Adaptive Tests for Improving Adaptivity

Unhee Ju, Riverside Insights; Mark Reckase, Psychometric Solutions; JongPil Kim, Riverside Insights

Although multistage testing has more practical advantages over full item-level adaptive testing, there is a tradeoff caused by its reduced level of adaptation. This study aims at investigating the performance of multiple objective functions using adaptation indices to assemble multistage tests for better adaptivity and measurement efficiency.

Comparison of On-the-fly MST with Preassembled MST on PISA Data

Xiuxiu Tang; Tong Wu; Yi Zheng, Arizona State University; Kit-Tai Hau, Chinese University of Hong Kong; Hua-Hua Chang, Purdue University

Recently, Multistage Adaptive Testing (MST) has been adopted by numerous large-scale testing/assessment programs including GRE and PISA. This study plans to compare the efficiency and accuracy of OMST with that of preassembled MST using the framework of PISA 2018 reading test.

The Adoption of Adaption: A State's Transition from Fixed Forms to MST

Anthony D. Fina, Iowa Testing Programs; Mingjia Ma, The University of Iowa

This paper examines the measurement challenges of a state's transition to an MST and compares simulated results with operational data. It also summarizes how practical considerations inform the final model, including item bank characteristics, proficiency determinations, and publishing constraints. Guidance for other programs considering similar designs is provided.

Discussant:

Christopher Runyon, NBME

156. Growth/Longitudinal Applications

Paper Session

1:15 to 2:15 pm PT

Westin San Diego Gaslamp: Del Mar

Participants:

Grade Inflation Continues to Grow in the Past Decade

Edgar Sanchez, ACT; Raeal J. Moore, ACT

We saw evidence of grade inflation before and after accounting for student and school characteristics. Often, grade inflation became apparent in 2020 and 2021, or the rate of inflation increased during those years. Attributing these changes to the pandemic is difficult. This study supports the documented HSGPA inflation across decades

Applying human-centered design to enhance growth score reports

Jinah Choi, Edmentum, INC.; Audra Kosh, Edmentum, Inc.; Catherine Oberle, Edmentum; Natalie OShea, Edmentum

Building on recent methodological contributions to human-centered learning analytics, this research documents and describes the process of redesigning student growth reports which utilize gain scores from an interim assessment. Involving stakeholders early in the design of these reports ensures that educators make and communicate valid interpretations of student growth.

Discussant:

Dongmei Li, ACT

157. Response Time Advancements

Paper Session
1:15 to 2:15 pm PT
Westin San Diego Gaslamp: La Jolla

Participants:

General Cognitive Diagnosis Model for Response Time
Yi Chen, Teachers College Columbia University; Yi Yang, Teachers College Columbia University; Sizheng Zhu, Teachers College Columbia University; Young-Sun Lee, Teachers College Columbia University

This study proposes a general cognitive diagnosis model for response times (CDM-RT). The idea of the model is to consider person parameters defined on fine-grained attributes to improve the model fit and skill diagnosis related to RTs. The model framework is specified under two motivation conditions and two treatments of attribute profiles. Both simulation study and real data analysis are conducted via a fully Bayesian approach with Markov Chain Monte Carlo (MCMC) method.

Computer-Based Testing in Pandemic Mode: What Response Times May Tell Us
Mengyao Zhang, National Conference of Bar Examiners; Kylie Gorney, University of Wisconsin-Madison

This study is intended to explore the usefulness of response times for assessing the impact of the pandemic on test-taking behaviors. Both quantitative methods and data visualization techniques are used. The findings could help address validity issues and offer empirical evidence for further research. Practical implications are also discussed.

A New Multiple-Group DCM Model Incorporating Response Times, and Visual Fixations
Kaiwen Man, University of Alabama; Stefanie A. Wind, University of Alabama; Joni M Lakin, University of Alabama; Peida Zhan

Technology-enhanced learning system (TELS) has drawn much attention in educational assessment recently. To gain behavioral inferences by properly modeling multi-source data collected from TELS, this study proposes an innovative multi-group diagnostic classification model as an extension of the existing one for jointly modeling item responses, response times, and visual gazes.

Discussant:
Abby Javurek, NWEA

158. Analysis of Multiple Choice Items

Paper Session
1:15 to 2:15 pm PT
Westin San Diego Gaslamp: Santa Fe

Participants:

Utilizing Item Response Choice in Bayesian Multidimensional Adaptive Testing
Catherine Elizabeth Mintz, University of Iowa; Jonathan Templin, University of Iowa

Item distractors contain information regarding examinee ability, which can possibly be leveraged in Multidimensional Adaptive Testing (MAT) to provide shorter, more accurate tests. This study examines the benefit of modeling item response choice in MAT by generalizing MAT components for use with nominal response data.

A New CDM Framework for Diagnosing Skills and Misconceptions for Multiple-Choice Data
Jimmy de la Torre, University of Hong Kong; Xue-Lan Qiu, University of Hong Kong

The existing CDMs for MC data are suboptimal because they ignore some potential diagnostic information in the distractors. This study proposes a CDM framework that harness the data more effectively by accommodating skills, incomplete knowledge, and misconceptions. Simulation studies were conducted to evaluate the new model's parameter recovery and performances.

A Protocol to Evaluate the Comparative Measurement Value of Technology-Enhanced Items
Sebastian Moncaleano, Boston College

This paper presents the development of a protocol to judge the comparative measurement value of technology-enhanced items relative to stem-equivalent multiple-choice items based on psychometric properties and utility ratings. The protocol is applied to two forms of an instrument comprising classification and rank-ordering drag-and-drop items and stem-equivalent multiple-choice items.

Discussant:
Michael C. Rodriguez, University of Minnesota

159. Reliability Topics

Paper Session
1:15 to 2:15 pm PT
Westin San Diego Gaslamp: Plaza

Participants:

Exploring Rater Classification Accuracy and Rater Measurement Precision in Sparse Mixed-Format Assessments
Wenjing Guo, University of Alabama; Stefanie A. Wind, University of Alabama

We conducted simulation studies to explore the impacts of different rating designs on rater severity estimation and rater classification accuracy (severe/lenient). The results suggest that the complete rating design produced highest rater classification accuracy and greatest rater measurement precision, followed by the spiral link design and the MC link design.

Is Agreement Enough? An Exploration of Interrater Reliability
Bryndle L Bottoms, University of North Carolina at Charlotte; Timothy Scott Holcomb, University of North Carolina at Charlotte; Richard Lambert, UNC Charlotte

This study proposes a process for identifying evaluators that need additional training, and compares model-based indicators of potentially problematic raters with indexes based on percent agreement. Percent agreement statistics alone masked specific areas of evaluator practice where improvements could enhance the reliability, validity, and fairness of teacher evaluation scores.

Estimating Test-Retest Reliability under Self-Selection Bias on the Duolingo English Test
Will Belzak, Duolingo; J.R. Lockwood, Duolingo

We estimate test-retest reliabilities of the Duolingo English Test while adjusting for self-selection of test takers with respect to the decision to repeat and when to repeat. We estimate a reliability of approximately 0.90 for the DET overall score, and subscore reliabilities ranging from 0.86 to 0.88.

Discussant:
Robert Thomas Furter, Physician Assistant Education Association

160. Advancing Contemporary Validity Theory and Practice: An Interactive Town Hall

Organized Discussion
2:30 to 4:00 pm PT
Westin San Diego Gaslamp: California Ballroom A

A moderated interactive discussion with a panel of experts intended to address the following questions: - What aspect of modern validity theory do you think most supports the practice of validation? – What aspect of modern validity theory do you think most hinders the practice of validation?

Session Organizer:
Gregory Cizek, University Of North Carolina

Moderator:
Jon S. Twing, Pearson

Presenters:
Gregory Cizek, University Of North Carolina
Suzanne Lane, University Of Pittsburgh
Scott Marion, National Center for the Improvement of Educational Assessment

161. Diagnostic Measurement in Action

Coordinated Paper Session
2:30 to 4:00 pm PT
Westin San Diego Gaslamp: California Ballroom B

Since the publication of the seminal book Diagnostic Measurement: Theory, Methods, and Applications (Rupp, Templin, Henson, 2010), there has been a wave of research on diagnostic measurement models. Much of this research, however, has focused on methodological advancements and empirical applications are scarce. This session, coordinated by the Diagnostic Measurement SIG, presents a selection of applied diagnostic measurement studies. More specifically, each study presented highlights the added practical and interpretational value of diagnostic measurement models relative to traditional measurement models and their utility in applied research studies. The session will conclude with commentary from an expert in diagnostic models and their application in applied research studies.

Session Organizer:
Matthew James Madison, University of Georgia

Moderators:
Yu Bao, James Madison University
Qianqian Pan, The University of Hong Kong

Participants:

The Hidden Loss of Creative Potential in Schools: Is Girls' Creative Potential Under-Identified in Schools?
Sue Hyeon Paek, University of Northern Colorado; Yu Bao, James Madison University

Assessing Hong Kong Students' Digital Literacy Using a Multiple-Group CDM: A Cross-Sectional Study
Qianru Liang, The University of Hong Kong; Jimmy de la Torre, University of Hong Kong; Nancy Law, University of Hong Kong

A Dual-Purpose Model for Estimating Ability and Misconceptions
Wenchao Ma, University of Alabama; Miguel A. Sorrel, Universidad Autónoma de Madrid; Yuan Ge, University of Alabama; Xiaoming Zhai, Stanford University

Dynamic Learning Models Within a Cognitive Diagnosis Framework to Quantify Heterogeneous Learning Effectiveness of Learning Materials
Yanyan Tan; Shiyu Wang

Discussant:
Matthew James Madison, University of Georgia

162. Innovative Methods

Paper Session
2:30 to 4:00 pm PT
Westin San Diego Gaslamp: California Ballroom C

Participants:

Parameter Estimation for Mixed-Format Tests with No-U-turn sampler Hamiltonian Monte Carlo Method
Jiawei Xiong, University of Georgia; Xinhui Xiong, Educational Testing Service; Allan Cohen, University of Georgia; Bowen Wang, University of Florida

The calibration of mixed-format tests requires mixed models. This study employs a No-U-Turn Sampler Hamiltonian Monte Carlo (NUTS-HMC) algorithm, to estimate both the person and item posterior distributions with mixed models. Results indicate the NUTS-HMC performs at least as accurate and much more efficient than a standard Monte Carlo algorithm.

Conditionally Unbiased Best Linear Predictors for Score Augmentation
Xiang Liu, Educational Testing Service; Matthew Johnson, ETS; Sandip Sinharay, Educational Testing Service

The best linear predictor (BLP) has been proposed to combine different types of information in estimating true scores. The BLP is biased for individual examinees when conditioned on the true score. In this paper, we propose a conditionally unbiased BLP. Additionally, a least square method is introduced for parameter estimation.

Model Selection for Latent Dirichlet Allocation in Assessment Data: New Advances
Constanza Mardones, University of Georgia; Hye-Jeong Choi, Human Resources Research Organization; Minju Hong, University of Georgia; Allan Cohen, University of Georgia

Perplexity and cross-validation are studied to determine their performance for model selection for latent Dirichlet allocation in practical testing conditions with constructed-response items. Preliminary results based on 10 replications suggest that perplexity might not accurately select the best fitting model in this context. A total of 30 replica- tions will be reported for the conference presentation. Keywords: Model selection, latent Dirichlet Allocation, Bayesian estimation.

Modeling audio passage "listenability" and the development of listening comprehension
Alistair Van Moere, MetaMetrics Inc; Jing Wei, MetaMetrics Inc; Michael J Fox, MetaMetrics

Listening is as essential to academic success as reading. However, until now, no formula has been developed to measure the "listenability" of audio. This study investigates the development of a framework that models audio passage difficulty and the development of student listening ability, for comparison with reading ability.

Statistical Software and Quantitative Methods in Measurement Research
Brandon LeBeau, University of Iowa; Yichong Cao; Ariel M. Aloe, University of Iowa

Quantitative research relies heavily on statistical tools and statistical procedures. This study looks at the current statistical software and methods in measurement research using research synthesis methods. This paper will explore the frequency of software citation and the use of statistical methods in published measurement research.

Discussant:
Susan Lottridge, Cambium Assessment, Inc

163. Assessment Issues in Competency-Based Education and Micro-Credentialing

Coordinated Paper Session
2:30 to 4:00 pm PT
Westin San Diego Gaslamp: Del Mar

Competency-based education involves the specification and organization of competencies (with one strategy being learning map systems), assessments, and recognizing achievements using micro-credentials. In such a system within higher education and professional learning, issues arise in the development and alignment of assessments along with the evidence needed to support their quality. In this session four presentations and a discussant will indicate the role of assessments in competency-based education programs and micro-credentialing, as well as issues and procedures to ensure their quality. The first presenter will indicate the psychometric, practical, and political challenges associated with many micro credentialing programs. The second presenter will use a validity framework to describe the assessment challenges facing competency-based micro credentials along with some potential solutions. The third presenter will indicate the processes and results of work undertaken on an online competency-based higher education to ensure the quality of the assessment system. The fourth presenter will describe the systems that need to be in place to realize an aligned, accurate, fair, and valid assessment system. Finally, a discussant will provide a perspective of assessment in competency-based education that leads to micro-credentials, as well as offer some comments about the presentations.

Session Organizer:
Thanos Patelis, Teachers College at Columbia University & University of Kansas

Chair:
Andrew Wiley, ACS Ventures

Participants:

Opportunities and Challenges in the Development of New Micro Credentialing Programs
Andrew Wiley, ACS Ventures

Competency Assessments for Micro-Credentialing: Building a Validity Argument
Katie McClarty, Renaissance

Ensuring Quality Assessment in Online Competency-Based Higher Education
Heather Hayes, Western Governors University

The Assessments in a Competency-Based Higher Education Program with Micro Credentials
Thanos Patelis, Teachers College at Columbia University & University of Kansas

Discussant:
Stephen G Sireci, University of Massachusetts, Amherst

164. Model-based Approach to Oral Reading Fluency Assessment

Coordinated Paper Session
2:30 to 4:00 pm PT
Westin San Diego Gaslamp: La Jolla

Oral reading fluency (ORF) has been regarded as an important indicator of overall reading competence and assessed mostly as part of curriculum-based measurements to screen at-risk readers. Also, there are many standardized tests that use ORF assessments. Meanwhile as part of a Response to Intervention (RTI) framework, ORF measures have been widely used as screening tools to help identify students at risk for poor achievement outcomes, and as progress monitoring tools to help teachers determine effective instruction and monitor students reading growth. As part of an effort to develop an improved ORF assessment system, an improved ORF assessment system has been developed. Also, a new model-based approach to ORF assessment scores has been developed that has enhanced some psychometric properties of ORF scores. Currently, an R package is being developed for a wider use of the model-based approach to ORF assessment. Additionally, more estimators of ORF scores have been derived and have been incorporated into the R package. This coordinated paper session will aim to demonstrate the theory behind the model-based approach to ORF, as well as some practical implications of this new approach.

Session Organizers:
Akihito Kamata, Southern Methodist University
Joseph F. T. Nese, University of Oregon

Participants:

Estimating Passage Parameters by the Model-based Approach to ORF Assessment
Cornelis Potgieter, Texas Christian University; Yusuf Kara, Southern Methodist University; Akihito Kamata, Southern Methodist University

Estimating Fluency Scores by the Model-based Approach to ORF Assessment
Akihito Kamata, Southern Methodist University; Cornelis Potgieter, Texas Christian University; Yusuf Kara, Southern Methodist University; Sarunya Somsong, Srinakharinwirot University & Southern Methodist University; Kuo Wang, Southern Methodist University

Evaluation of Various Estimators for Model-based Fluency Scores
Sarunya Somsong, Srinakharinwirot University & Southern Methodist University; Kuo Wang, Southern Methodist University; Anh Thu Le, Southern Methodist University; Yusuf Kara, Southern Methodist University

Practical Implications of the Model-based Approach to ORF Assessment
Yusuf Kara, Southern Methodist University; Joseph F. T. Nese, University of Oregon; Akihito Kamata, Southern Methodist University

Discussants:
Hong Jiao, University of Maryland
Yaacov Petscher, Florida Center for Reading Research

165. Analysis of Polytomous Item

Paper Session
2:30 to 4:00 pm PT
Westin San Diego Gaslamp: Santa Fe

Participants:

An Explanatory Mixture IRT Model for Careless Responding in Survey Data

Esther Ulitzsch; Seyma N. Yildirim-Erbasli, University of Alberta; Guher Gorgun, University of Alberta; Okan Bulut, University of Alberta

An explanatory mixture IRT model for questionnaire data is presented that identifies careless responses on the respondent-by-item level. Its utility for investigating person and item characteristics associated with careless responding is illustrated on Big Five inventory data.

Multilevel Modeling of Item Parameter Drift in Polytomous Items

Wei Xu, American Board of Anesthesiology; William J Muntean, National Council of State Boards of Nursing

The presence of IPD might significantly threaten test score validity. Most studies examined IPD effect on dichotomous items but paid relatively limited attention to IPD effect on polytomous items. The proposed study used HGLM to investigate IPD on polytomous items. Simulation results suggested the proposed HGLM model is effective.

Number of response categories and sample size requirements in polytomous IRT models

Dubravka Svetina Valdivia, Indiana University; Shenghai Dai, Washington State University

Applications of polytomous IRT Models in applied fields (e.g., health, education, psychology) are abound. However, little is known about the impact of the number of categories and sample size requirements for precise parameter recovery. Our simulation study investigates these issues to inform applied researchers for survey development.

Towards Quantifying Model Complexity in Polytomous Item Response Models

Yon Soo Suh, UCLA; Li Cai, UCLA

This study introduces an innovative method for generating the complete categorical data space and corresponding modeling approaches that allows investigations into the fitting propensity of a variety of item response models (IRMs). We evaluate the performance of the proposed approach in the context of polytomous IRMs consisting of many items.

Discussant:

Frank Rijmen, Cambium Assessment, Inc

166. Advances in DIF

Paper Session
2:30 to 4:00 pm PT
Westin San Diego Gaslamp: Plaza

Participants:

A Comparison of the Anchor Selection Strategies for DIF Analysis

Haeju Lee, University of North Carolina Greenboro; Kyung Yong Kim, University of North Carolina Greenboro

In this study, we investigate the performance of the three anchor selection procedures (Wald test-AATA, IRT-LRT AOAA, and IRT-LRT forward procedures) paired with two criteria (MinG^2/Min▯^2 and MaxA) by calculating the accuracy of the designed anchor items and the follow-up DIF tests' power and Type ▯ error rates.

An IRT-Based Geometrical Approach to Differential Item Functioning

Leah Feuerstahler, Fordham University

Differential item functioning (DIF) assessment in the context of item response theory (IRT) requires identifying an anchor measure that does not differ between groups. This study proposes a measure of DIF based on a geometrical interpretation of the IRT metric. This new measure is illustrated in real and simulated data.

DIF without anchors: A comparison of some procedures from robust regression

Peter Halpin, UNC-Chapel Hill

DIF with respect to a categorical variable can be framed as a problem in robust regression. This leads to a discussion of how concepts and techniques from robust statistics can shed light on some long-standing issues in DIF analysis, notably how to assess DIF without anchor items.

Detecting DIF associated with Multiple Covariates in Multidimensional 2PL Models

Ruoyi Zhu, University of Washington; Chun Wang, University of Washington

A new family of DIF detection methods, namely, the regularization methods, have the advantage of handling multiple correlated covariates simultaneously. We explore two regularization methods---Lasso and adaptive Lasso---in detecting DIF associated with multiple covariates in Multidimensional 2PL Models. Simulation studies will be presented for illustration.

Discussant:

***Benjamin R. Shear*,** University of Colorado Boulder

167. NCME Board of Directors Meeting #2 (Invite Only)

4:00 to 7:00 pm PT
Westin San Diego Gaslamp: Imperial

Participant Index

A

NAME	SESSION NUMBER
Abad, Francisco J.	032, 069
Abdalla, Widad	154
Abulela, Mohammed	127
Adams, Elizabeth L.	106, 110
Ahmed, Ishita	097
Akhuseyinoglu, Kamil	113
Albano, Anthony	118, 120, 149
Albert, Alexi	114
Allen, Jeff M.	079
Allen, Jessica	110
Almond, Russell G	049, 145
Aloe, Ariel M.	162
Alvero, A J	075
An, Lily	067, 120, 133
An, Xiaozhu	130
Andersson, Bjorn	108
Andrade, Heidi	105
Anghel, Ella	035
Ankomah, Francis	122
Ann-Kathrin, Bültmann	100
Antal, Judit	114
Antonio, Anthony	075
Araneda, Sergio	091, 143
Arce, Alvaro J.	112
Arcilla, Mary Anne	141
Armour-Thomas, Eleanor	066
Armstrong, Maria	141
Arslan, Burcu	070
Arthur, Ann	106, 130
Arthur, David	019
Attali, Yigal	152
Austin, Bruce	130
Awwal, Nafisa	100

B

NAME	SESSION NUMBER
Bai, Fangxing	032
Bai, Yifan	076, 153
Bailey, Alison	105
Bailey, Paul	051, 063, 094
Baker, Eva L.	066
Baldwin, Peter	121, 142, 152
Ballard, Laura D	151
Banjanovic, Erin	092
Bao, Yu	144, 161
Barbosa, Denilson	113
Barnes, Sophie	067
Barton, Reka	128
Basaraba, Deni	026
Bashkov, Bozhidar M.	126, 130
Bastedo, Michael	075
Bausch, Emma	075
Becker, Kirk	005, 013, 099, 142
Becker, Rachael N	110
Bei, Ni	072
Beiting-Parrish, Magdalen	091, 114
Belzak, Will	159
Bennett, Randy	066, 146
Bergner, Yoav	106
Bertling, Masha	097
Betebenner, Damian	040, 043, 061, 119, 132
Beverly, Tanesia	019, 135
Bezirhan, Ummugul	065, 122
Bhaktha, Nivedita	034
Bichler, Sarah	011
Binici, Salih	094
Bishop, Kyoungwon Lee	106, 114
Bochenek, Jennifer	075
Bohrnstedt, George William	076
Bolt, Daniel	010, 079, 135, 144, 148
Bonett, John	007
Booth, Brandon M.	110
Borglum, Joshua	142
Borgonovi, Francesca	108, 129
Bottoms, Bryndle L	159
Boughton, Keith	109
Bovaird, James	106
Bowen, Daniel	006
Bowzer, Angela	018
Boyd, Aimee	017
Boyd, Brittany	076
Boyer, Michelle	069, 089, 122
Boykin, Allison Ames	050, 106, 131, 140
Bradford, Allison	011
Bradshaw, Laine	107, 144
Brandt, Chris	067
Brennan, Robert	059, 118, 142
Brice, Amanda	058
Bridgeman, Brent	060
Briggs, Derek	088, 101, 133, 136, 146
Brodersen, Alex	079
Broer, Markus	153
Brown, Terran,	080
Brumfield, Adele	112
Buehler, Eric	051
Bulut, Hatice Cigdem	094
Bulut, Okan	062, 094, 113, 122, 135, 165
Bunch, Michael Brannen	006, 059, 108
Bynum, Bethany	092
Byun, Bo-Kyung	075

C

NAME	SESSION NUMBER
Cai, Li	030, 082, 165
Cai, Liuhan	004, 017, 113, 144
Caines Lee, Jade	128
Caliço, Tiago A.	027, 044, 116

Participant Index

Camara, Wayne J.	008, 016, 125
Camargo Salamanca, Sandra Liliana,	122
Camilli, Gregory	008
Cancado, Luciana	133
Cao, Yichong	162
Cappaert, Kevin	062, 133
Carr, Peggy	104
Casabianca-Marshall, Jodi	150
Casillas, Alex	120
Cayton-Hodges, Gabrielle	129
Chakraborty, Roti	126
Chalmers, R. Phillip	019
Chang, Hua-Hua	019, 131, 155
Chao, Hsiu-Yi	034
Chatterji, Madhabi	066
Chavez, Carlos	130
Chen, Fu	035, 126
Chen, Haiqin	017
Chen, Jianshen	011
Chen, Jihang	131
Chen, Jing	063, 130
Chen, Jyun-Hong	034
Chen, Ping	012
Chen, Siqi	018, 025
Chen, Tavy	018
Chen, Wenya	111
Chen, Xiaoyu	111
Chen, Yi	135, 157
Cheng, Yiling	114
Cheng, Ying	064, 069, 098, 126, 152
Chien, Yuehmei	003
Chiu, Chia-Yi	032, 039, 042
Cho, Sun-Joo	012
Cho, YoungWoo	130
Cho-Baker, Sugene	075
Choe, Edison M.	086
Choi, Hye-Jeong	092, 162
Choi, Jinah	100, 156
Choi, Kilchan	130
Choi, Seung W.	002, 068, 116
Circi, Ruhan	044, 070, 116, 123, 130
Cisterna-Alburquerque, Dante	073
Cizek, Gregory	115, 154, 160
Clark, Amy	030, 107, 143
Clark, Corey	106
Clauser, Amanda	072
Clauser, Brian	059, 088, 142, 149
Clements, Douglas	106
Coggeshall, Whitney	084
Cohen, Allan	070, 162
Cohen, Yoav	150
Collier, Tina	017
Connally, Mark R	118
Cook, Linda	021, 125
Cook, Robert	028
Cooperman, Allie	040, 043, 132
Cousar, Davis	076
Crisp, Dianne	106
Cruce, Ty	120
Cuddy, Monica	072
Cui, Ying	035, 126
Custer, Michael	114
Cutumisu, Maria	126

D

NAME	SESSION NUMBER
Dadey, Nathan	011, 040, 043, 132
Dai, Shenghai	106, 122, 130, 165
Dailey, David	114
Daisher, Ted	122
Dallas, Andrew	027, 072
Darling-Hammong, Linda	104
Davidson, Anne H.	021
Davis, Jennifer	099
Davis, Laurie	101, 117, 133
Davis-Becker, Susan	018, 028, 087, 151
Dawson, Matt Matt	133
Day-Hess, Crystal	106
Deane, Paul	145
De Boeck, Paul	031, 071
de la Torre, Jimmy	032, 036, 055, 069, 158, 161
Demir, Cihan	114
Demirkaya, Onur	010, 106, 122
Demmans Epp, Carrie	113
Denbleyker, John	079
Deng, Jiayi	019, 077, 127
Depaoli, Sarah	106
Deribo, Tobias	129
Diao, Qi	007, 071
Diaz, Ted	023, 092
Ding, Yishan	010, 098
D'Mello, Sidney	110
Do, Tai	093
Domingue, Ben	075, 091, 097, 149
Dong, Yixiao	106
Dorans, Neil	033
Dore, Kelly	009
Dozier, Sara	064
Drost, Bryan R.	024
Du, Yang	131
Du, Ying	147
Duan, Qizhou	098
Dumas, Denis	106, 148
Dunbar, Stephen B.	151

E

NAME	SESSION NUMBER
Edi, Daniel	082
Egan, Karla	020, 022, 109
Embretson, Susan	084, 122, 130
Engelhard, George	025
Englert, Kerry	110

Participant Index

Ercikan, Kadriye	074, 104
Ersan, Ozge	035
Escalante, Sean Oliver	032
Everson, Howard	066, 091
Ezike, Nnamdi	106, 131, 150
Ezzelle, Carol	009

F

NAME	SESSION NUMBER
Fahle, Erin Michelle	149
Fairchild, Amanda	114
Famularo, Lisa	144
Fan, Fen	072, 130
Faulkner-Bond, Molly	096
Fayaz, Tabasom	106
Fechter, Tia	013, 023
Feinberg, Rich	045, 047, 113
Feng, Xiaoying	031, 094, 122
Ferrara, Steve	028, 030, 074
Feuerstahler, Leah	100, 118, 121, 166
Filonczuk, Audrey	064
Fina, Anthony D.	155
Finch, Holmes	114
Fink, Aron	069
Finn, Bridgid	070
Finney, Sara	010, 091
Firoozi, Tahereh	113, 149
Fishbein, Bethany	065
Fisk, Charles	114
Fitzpatrick, Joseph	106
Fitzpatrick, Steven	036
Flores, Charity	061
Forte, Ellen	058
Forzani, Elena	076
Foster, David	124
Foster, Paul	106
Fox, Michael J	162
Foy, Pierre	065
Fraillon, Julian	076
Francis, Catherine Xueying	079
Fremer, John	085
French, Brian	106, 114, 122, 130
Frey, Andreas	069
Frey, Sharon	018
Fu, Yanyan	078
Fujimoto, Ken	111
Furgol Castellano, Katherine	024, 067, 119
Furter, Robert Thomas	013, 159

G

NAME	SESSION NUMBER
Gao, Ruiyan	034
Gao, Yizhu	122
Garcia, Elda	115, 141
Gauran, Iris Ivy	032
Ge, Yuan	121, 161
Gebre-Medhin, Ben	075
Gee, Jim	104
Geisinger, Kurt F	059
Geranpayeh, Ardeshir	018
Gerasimova, Daria	130
Gholson, Melissa L.	021, 115
Gianopulos, Garron	030, 063, 068
Giebel, Sonia	075
Gierl, Mark	124, 149
Gilbar, Charlotte	105
Gochyyev, Perman	106
Goldhammer, Frank	031, 129
Golovkina, Olga	092
Gong, Brian	030
Gong, Tao	057
Gonulates, Emre	092
Gonzalez, Katrina	119
Goodman, Joshua	013, 027, 149
Goodwin, Amanda	012
Gorgun, Guher	062, 077, 122, 165
Gorney, Kylie	078, 157
Gotch, Chad	130
Gotwals, Amelia	102
Gow, Andrew	114
Grabovsky, Irina	085
Grady, Matt	017
Greiff, Samuel	084
Gridiron, Tina	120
Grover, Raman	080
Groves Price, Paula	096
Gu, Lixiong	114
Guo, Hongwen	007, 024, 046, 108
Guo, Wenjing	144, 159

H

NAME	SESSION NUMBER
Habing, Brian	114
Hahnel, Carolin	031
Haider, Muhammad Qadeer	070
Halpin, Peter	082, 152, 166
Hamilton, Laura	073, 139
Han, Kyung (Chris)	053, 056, 078, 086, 098
Han, Suhwa	149
Han, Zhuangzhuang	007
Hansen, Mark	151
Hao, Jiangang	024, 038, 041
Haring, Samuel	116

Participant Index

Harring, Jeffrey	031
Harris, Deborah	062, 112
Harris, Michael Wayne	114
Harris, William	120
Hastedt, Dirk	076
Hatting-Niekerk, Ashley Kevin	036
Hau, Kit-Tai	155
Hauenstein, Clifford Erhardt	122
Hayes, Heather	163
He, Qiwei	057, 094, 129
He, Wei	020
He, Yi	017
He, Yong	086, 134
Heck, Tamara	073
Hembry, Ian	143
Hembry, Tracey	029
Hemphill, Cadelle	027
Henderson, Diane	112, 120
Henson, Robert	144
Heritage, Margaret	102
Hermann, Pamela	109
Hicks, Juanita	027, 044, 116, 123
Himelfarb, Igor	082, 114
Himmelsbach, Zach	120
Hinds, Fiona	120
Hjalmarson, Margret	130
Ho, Andrew	067, 075, 090, 117, 133
Ho, TsungHan	106
Hochweber, Jan	131
Hodgin, Erica	073
Hoeve, Karen	147
Holcomb, Timothy Scott	159
Hong, Hyeri	026, 111
Hong, Maxwell	064, 126
Hong, Minju	162
Hooper, Martin	076
Hoover, Jeffrey	121, 153
Houck, Brooke	010, 114
Hu, Ann	020, 035, 078
Hu, Mingqi	079
Hua, Cheng	052
Huang, Qi	144
Huang, Yingshi	012
Huang, Yue	113
Hudson, Kim	022
Huff, Kristen	096
Huggins-Manley, Anne Corinne	110, 140, 152
Hughes, Gerunda	066, 128
Hunter, Charles Vincent	114

I

NAME	SESSION NUMBER
Ihlenfeldt, Samuel Dale	077, 127
Ingrisone, James	010, 033
Ingrisone, Soo	010, 017
Irribarra, David Torres	088
Ivanova, Militsa	036
Iwatani, Emi	146

J

NAME	SESSION NUMBER
Jackson, Janine	122
Jackson, Tanner	139
Jafari, Amir	048
Jappie, Naziema	036
Javurek, Abby	157
Jewsbury, Paul Adrian	007
Ji, Xuejun Ryan	012
Jia, Li	014
Jiang, Bingnan	116
Jiang, Yang	129
Jiao, Hong	098, 111, 164
Jimenez, Auburn	114
Jin, Kuan-Yu	025
Jin, Rong	018
Jin, Ying	092, 114
Jing, Shumin	134
Johnson, Chanda	107
Johnson, David	123
Johnson, Mark	093
Johnson, Matthew	067, 135, 150, 162
Johnston, Mary M.	147
Joldersma, Kevin	147
Jolin, Jerred	106
Jones, Andrew	013
Jones, Edmund	018
Jones, Eli Andrew	121
Jones, Paul Edward	099, 142
Jonson, Jessica L.	008, 026
Joo, Sean	071, 082, 108
Jose Klinowski Gomez, David	149
Ju, Unhee	018, 082, 155
Julian, Marc W	022, 109
Jung, Hyun Joo	114
Jung, Ji Yoon	065
Jurich, Daniel	130
Juster, Fern	009
Justice, Derek	134

K

NAME	SESSION NUMBER
Kahne, Joseph	073
Kamata, Akihito	083, 164
Kanaya, Tomoe	128
Kane, Michael	009
Kanistra, Voula	028
Kannan, Priya	045
Kanopka, Klint	145

Kao, Jenny	151
Kao, Shu-chuan	019
Kaplan, David	065
Kapoor, Radhika	149
Kapoor, Shalini	130
Kara, Yusuf	083, 164
Karamese, Hacer	015
Karvonen, Meagan	107
Kaufman, Julia	073
Kehinde, Olasunkanmi	122
Kelbaugh, Chris	007
Kelberlau, Darin	105
Kelcey, Benjamin	032
Keller, Lisa	085, 114, 130, 149
Keng, Leslie	035, 067
Kern, Justin L.	114
Ketterlin Geller, Leanne	070, 106
Kettler, Ryan	008
Keum, Eunhee	130
Khorramdel, Lale	065
Kim, Dong-In	022, 109, 114
Kim, Eunbee	130
Kim, Han Yi	004
Kim, Hyung Jin	118, 122
Kim, JongPil	018, 082, 155
Kim, Jungnam	086
Kim, Kyung Yong	122, 166
Kim, Nana	135
Kim, Se-Kang	001, 130
Kim, Seongeun	072, 122
Kim, Sooyeon	020, 142, 150
Kim, Stella	012, 033, 083, 118, 122
Kim, Sunhee	020
Kim, Yeonwho	012
Kim, Yongnam	020
Kim, YoungKoung	106
Kim, Young Yee	031, 094, 122
Kim, Yun-Kyung	021, 030, 082, 130
King, David	113
King, Teresa	075
Kingsbury, G. Gage	127
Kintz, Tara	102
Kitmitto, Sami	076
Klinger, Corrie Rebecca	126
Klinkert, Lawrence J.	106
Klugman, Emma M.	075, 120
Knezevich, Lily	009
Kobrin, Jennifer	060, 121, 143
Koc, Piotr	108
Koehn, Hans Friedrich	039, 042
Kollias, Charalambos	011
Kolovou, Dimitra	131
König, Christoph	069
Kosh, Audra	100, 117, 156
Kreitchmann, Rodrigo Schames	032, 069
Kroehne, Ulf	031, 129
Kroopnick, Marc	092, 114
Kuhfeld, Megan	090, 117, 148
Kukea Shultz, Pohai	080, 110, 115
Kuklick, Livia	084
Kumar, Lavanya Shravan	082
Kyllonen, Patrick Charles	071

L

NAME	SESSION NUMBER
Ladaga, Bea Margarita	014, 025
LaFlair, Geoff	114
Lai, Hollis	124
Lakin, Joni M	157
Lam, Patrick	106
Lambert, Richard	159
Lamm, Rik	093
Landl, Erika	105
Lane, John	102
Lane, Suzanne	009, 029, 074, 124, 160
Langi, Meredith	079
Larson, Eric C.	106
Law, Nancy	036, 161
Le, Anh Thu	164
LeBeau, Brandon	162
Lee, Dukjae	072
Lee, Haeju	122, 166
Lee, Hollylynne	144
Lee, Hyeryung	026, 111
Lee, Jonghwan	068
Lee, Michael	051, 116
Lee, Mina	108
Lee, Seo Young	020
Lee, Sooyong	068
Lee, Sunhyoung	106
Lee, Won-Chan	012, 015, 118, 122
Lee, Young-Sun	135, 157
Lee-Robbins, Elsie	027
Leighton, Jacqueline P.	126
Leite, Walter	152
Leventhal, Brian C	050, 077, 091, 125, 131
Lewis, Daniel	028, 058, 068
Li, Anqi	011
Li, Chen	007, 024
Li, Dongmei	106, 149, 156
Li, Hongli	016, 114, 126
Li, Jie	017, 019
Li, Jie	018
Li, Jun	016
Li, Jun	098
Li, Scott	100
Li, Shirley	018
Li, Wen-Ching	022
Li, Xin	124
Li, Xin	130
Li, Xueming	035, 078
Li, Zhushan Mandy	131
Li, Ziying	152

Participant Index

Liang, Longjuan 079
Liang, Min 012
Liang, Qianru 036, 161
Liao, Dandan 018, 121
Liao, Xiangyi 079
Liaw, Yuan-Ling 070
Lim, Hwanggyu 086
Lim, Sangdon 068
Lim, Youn Seon 014, 032
Lin, Chih-Kai (Cary) 026
Lin, Jie 025
Lin, Qiao 020
Lin, Ye 020
Lin, Zhongtian 019, 079
Lindner, Marlit Annalena 084
Lindner, Marlit 077, 082
Ling, Guangming 075
Litschwartz, Sophie 097, 120
Liu, Cheng 069, 098
Liu, Chunyan 121, 130
Liu, Haiyan 035
Liu, Hongyun 031, 062
Liu, Huan 012, 122
Liu, Jin 144
Liu, Jinghua 142
Liu, Junhui 114
Liu, Lei 007, 073
Liu, Liu 017
Liu, Ou Lydia 075
Liu, Ren 035, 114
Liu, Xiang 046, 135, 162
Liu, Xiaowen 015
Liu, Yang 031, 063
Liu, Yue 031
Livne, Oren 038, 041, 071
Locke, Victoria 109
Lockwood, J.R. 119, 152, 159
Loken, Eric Eric 015, 135, 152
Lorie, William A. 035
Lottridge, Susan 048, 074, 089, 134, 162
Lu, Jing 098
Lu, Ru 007, 114
Lu, Yang 134
Lu, Yikai 069, 152
Lu, Ying 114
Luecht, Richard Melvin 059, 068, 101
Luedtke, Oliver 129
Luo, Yachen 094
Luo, Yong 054, 079
Lutsyk, Alina 035, 126
Lyons, Susan 067, 110, 120, 128
Lyu, Weicong 135

M

NAME SESSION NUMBER

Ma, Chenchen 063
Ma, Mingjia 122, 155
Ma, Wenchao 055, 106, 161
Ma, Ye 026, 062
MacPherson, Anna 064
Madison, Matthew James 079, 161
Malatesta, Jaime 033, 118
Man, Kaiwen 052, 157
Manna, Venessa Q 007, 080
Mao, Xia 078
Mardones, Constanza 162
Marion, Scott 067, 107, 139, 154, 160
Marks, Gary Neil 108
Martin, Michael 065
Martineau, Joseph A. 021
Martinez Orbegozo, Eva Flavia 106
Masters, Jessica 144
Mattern, Krista 120
Maul, Andrew 081
McCaffrey, Daniel 067, 074, 119
McCall, Martha 143
McClarty, Katie 117, 163
McCluskey, Sydne 091, 113
McCormick, Carina M. 100
McDevitt, Matthew 075
McGrane, Joshua 088
McIntee, Kimberly G. L. 016
McKinley, Danette Waller 009
McMurrer, Jennifer 070
Mee, Janet 149
Mejia, Ivy 014, 032
Meng, Huijuan 033, 099
Mertens, Ute 082
Meyer, Patrick 090
Miao, Jing 150
Michaelides, Michalis 036
Miller, Angela 130
Minchuk, Ilona 129
Mintz, Catherine Elizabeth 158
Mireles, Nicolas 144
Mislevy, Robert J. 081
Mix, Daniel 101
Mo, Ya 114
Moloney, Danielle 110
Moncaleano, Sebastian 131, 158
Montgomery, Melinda 086
Moore, Raeal J. 156
Morales, Amy 149
Morell, Linda 064
Morrison, Kristin M. 062, 133
Morse, Kathleen 126
Moses, Tim 033, 106, 142
Muckle, Timothy 085
Mueller, Lorin 130
Mundhenk, Kimberly 119

Participant Index

Muntean, William J	019, 165
Murchan, Damian	093
Murphy, Daniel	018
Musturi, M Hassan	113
Muszyński, Marek	087
Myers, Aaron	092

N

NAME	SESSION NUMBER
Nabors Olah, Leslie	129
Naeim Abadi, Ali	113
Nagy, Gabriel	077
Nájera, Pablo	032, 039, 042, 069
Naumann, Alexander	131
Naumann, Johannes	031
Naveiras, Matthew David	012
Nazari, Sanaz	149
Nese, Joseph F. T.	083, 164
Nichols, Paul	068
Nicola, Tara P.	120
Niu, Chunling Chunling	153
Niu, Luping	068

O

NAME	SESSION NUMBER
Ober, Teresa	069, 126
Oberle, Catherine	156
O'Donnell, Francis	072, 113
Ogut, Burhan	123, 130
Oh, Hyeon-Joo	056
Olea, Joemari	014, 025, 032
Olivera Aguilar, Margarita	080
O'Neill, Thomas	147
Ong, Thai Quang	047, 113
O'Riordan, Maura	012
Ormerod, Christopher	048
Orona, Gabe Avakian	087
O'Rourke, Kevin	114
Ortiz, Samuel O.	008
Osborne, Jonathan Francis	064
OShea, Natalie	156
Ouyang, Wenli	121
Ouyang, Xiangzi	123
Ozer Ozkan, Yesim	126

P

NAME	SESSION NUMBER
Pace, Jesse R.	114
Pace, Lillian	029
Padminiamma, Nisha	086
Paek, Insu	019, 130
Paek, Sue Hyeon	161
Paino, Maggie	061
Palermo, Corey	074, 089, 134
Palma, Jose R.	114
Pan, Qianqian	106, 161
Pandian, Ravi	072
Paniagua, Miguel	149
Park, Bitnara Jasmine	076
Park, Minjeong	031
Park, Seohee	111, 122
Park, Yoon Soo	020
Parshall, Cynthia	013
Pastor, Dena	010, 077
Patarapichayatham, Chalie	109
Patelis, Thanos	060, 163
Patz, Richard	123
Pavlakis, Alexandra E	080
Peabody, Michael R.	062, 085
Pearman, Alvin	075
Pearson, P David	104
Peck, Frederick	101
Pedrajita, Jose	014
Pellegrino, James	107
Peng, Fang	005, 019
Perie, Marianne	029
Perkins, Beth	010
Petscher, Yaacov	164
Pham, Duy N.	079
Phenow, Aurore Yang	022, 109
Pitoniak, Mary	125
Plackner, Christie	022
Plourde, Jessica	118
Podrouzek, Wayne	106
Poggio, John	120
Pokropek, Artur	087, 108
Pommerich, Mary	092
Poppen, Marcus	130
Por, Han-Hui	007
Potgieter, Cornelis	083, 164
Potter, Andrew	113
Powers, Sonya	018
Praetorius, Anna-Katharina	131
Primoli, Vince	142
Proctor, Thomas	011
Pruitt-Britton, Tiffini	110
Puhan, Gautam	007, 142
Pullen, Kendra	105

Participant Index

Q

NAME	SESSION NUMBER
Qiao, Xin	111
Qiao, Xin	054
Qiu, Xue-Lan	158
Qu, Yanxuan	019
Quansah, Frank	122

R

NAME	SESSION NUMBER
Raadt, Jay Schyler	080
Rabe-Hesketh, Sophia	130
Rabinowitz, Stanley N	100
Rajeb, Mehdi	052
Randall, Jennifer	059
Raymond, Mark	009
Rebouças-Ju, Daniella	098
Reckase, Mark	023, 114, 155
Reiter, Harold	009
Ren, He	012
Rewley, Kelly	016
Reynolds, Katherine	065
Ribeiro, Ana Carolina Trindade	149
Richards III, Randall	110
Rickels, Heather	151
Rijmen, Frank	018, 121, 154, 165
Riordan, Brian	011
Rios, Joseph A.	019, 108, 112, 127, 140
Riveros Medelius, Nicolas	106
Robin, Frederic	108
Rodrigues, Keri	141
Rodriguez, Michael C.	016, 035, 093, 096, 130, 146, 158
Roduta Roberts, Mary	130
Roeber, Edward Dean	102
Rome, Logan	117, 133
Roumaya, Matt	047
Roussos, Louis	004, 017, 113, 144
Ruiz-Primo, Maria Araceli	104
Runyon, Christopher	114, 155
Rusticus, Shayna	106
Rutkowski, Leslie	108
Ryoo, Hyun Suk	122
Ryoo, Ji Hoon	122

S

NAME	SESSION NUMBER
Sahin, Fusun	076
Sahin, Sakine Gocer	106, 114
Salas, Jorge	012
Sanchez, Edgar	156
Sanei, Hamid	144
Santos, Kevin Carl	014, 025, 032
Sarac, Merve	113
Sarama, Julie	106
Sarraf, Shimon	130
Satkus, Paulius	077
Sato, Edynn	021, 030, 130, 151
Sauder, Derek	092
Schellman, Madeline	144
Schmidt, Susanne	100
Schneider, Christina	063, 068, 116, 130
Schneider, Wei S.	011, 020, 036
Schonberg, Christina	130
Schweig, Jon	148
Schweig, Jonathan	090
Secolsky, Charles	110
Sedransk, Nell	114
Segall, Daniel	023, 092
Seo, Daeryong	130
Sharairi, Sid	018
Shear, Benjamin R.	093, 166
Shelton, Stephanie	125
Shenavai, Kevin	100
Shi, Dexin	114
Shi, Qingzhou	052
Shin, David	003, 017, 025, 036
Shin, Hyo Jeong	108
Shin, Nami	130, 151
Shivraj, Pooja	072
Sicangco, Emir Lenard Suerte Felipe	034
Siddiq, Fazilat	093
Sikali, Emmanuel	044, 051, 116
Silva, Monica	112
Silver, Daniel	067, 139
Sinharay, Sandip	019, 024, 098, 135, 150, 162
Sireci, Stephen G.	066, 112, 120, 141, 163
Sireno, Lisa	141
Skorupski, William	106, 130
Smith, Jessalyn	100
Smith, Russell	099
Soland, James	090, 108, 127, 148
Solano-Flores, Guillermo	104
Somsong, Sarunya	164
Song, Hao	085
Song, Yi	073, 145
Sorrel, Miguel A.	032, 039, 042, 069, 161
Stark, Stephen	082
Steedle, Jeffrey	077, 113, 118, 130
Steimel, Kenneth	011
Stevens, Mitchell	075
Stoeger, Jordan Nelson	015
Stone, Alexandra Lane	152
Stone, Alexandra	067
Stout, William	144
Strain-Seymour, Ellen	076
Strand, Paul	130
Strazzeri, Marian	025
Struthers, Vince	022
Student, Sanford	101
Su, Yu	017
Su, Yu	062

Participant Index

Wei, Xin	123
Weiner, John	124
Weir, J. B.	005, 149
Weissman, Alexander	135
Welch, Catherine	151
Wellberg, Sarah	101
Westine, Carl	122
Wheeler, Jordan M.	070
Wheeler, Kelley	151
Wibowo, Arianto	134
Wiley, Andrew	163
Wilhelm, Anne G.	110
Williamson, David	089
Wilmurth, Gina	078
Wilson, Jonee	110
Wilson, Joshua	113
Wilson, Mark	064, 088, 100, 106
Wind, Stefanie A.	121, 157, 159
Winfield, Angela	009
Winter, Sonja D	106
Wise, Steven	060, 077, 127
Wolf, Mikyung Kim	151
Wolfe, Edward	114
Wollack, James	078, 099
Wu, Amery	012, 031
Wu, Tong	122
Wu, Tong	155
Wu, Yi-Chen	123
Wu, Zebing	122
Wylie, Caroline	105
Wylie, E. Caroline	073
Wyse, Adam E	015, 031

X

NAME	SESSION NUMBER
Xiao, Jiaying	063, 085, 153
Xiao, Xingyao	130
Xiong, Jiawei	162
Xiong, Xinhui	162
Xiong, Yao	017, 018, 025, 036
Xu, Gongjun	063
Xu, Guanlan	026, 111
Xu, Jing	018
Xu, Shuangshuang	018
Xu, Wei	130, 165

Y

NAME	SESSION NUMBER
Yan, Duanli	049, 053, 145
Yan, Yan	072
Yaneva, Victoria	149
Yanez-Castillo, Natalia	110
Yang, Ji Seung	010, 025
Yang, Yi	122, 094, 135, 157
Yavuz, Sinan	094, 106, 153
Yee, Darrick	123, 130
Yildirim-Erbasli, Seyma N.	077, 122, 135, 165
Yin, Michelle	123
Yoo, Hanwook	056, 151
Yuan, Ye	025

Z

NAME	SESSION NUMBER
Zapata-Rivera, Diego	045, 049
Zenisky, April	045
Zhai, Xiaoming	161
Zhan, Pedia	157
Zhang, Haobai	031
Zhang, Jihong	123
Zhang, Jinming	010, 026, 106
Zhang, Jiwei	098
Zhang, Liru	134
Zhang, Maoxin	108
Zhang, Mengyao	033, 118, 157
Zhang, Mo	024, 046, 145
Zhang, Qiusi	123
Zhang, Susu	011, 106, 123, 131, 144
Zhang, Ting	051, 063, 094
Zhang, Xiao	123
Zhang, Yu	130
Zhang, Yuxiao	106
Zhang, Zhonghua	100
Zhao, Qian	016
Zhao, Yu (Tracy)	017, 036
Zhen, Mao Xiu	014
Zheng, Mingying	114
Zheng, Xiaying	031, 094, 122, 153
Zheng, Yi	140, 155
Zhou, Todd	098, 111
Zhou, Xuechun	130
Zhou, Ziwei	113
Zhu, Ruoyi	063, 086, 166
Zhu, Sizheng	135, 157
Zlatkin-Troitschanskaia, Olga	100
Żółtak, Tomasz	087
Zopluoglu, Cengiz	099
Zu, Jiyun	114

Participant Index

Suarez-Alvarez, Javier	129
Subhiyah, Raja G	121
Suh, Hongwook	086, 106
Suh, Yon Soo	079, 165
Suksiri, Weeraphat	064
Sun, Chuan	020
Sun, Guo-Wei	025
Sun, Huaping	072
Sun, Ting	083
Sussman, Joshua	100
Svetina Valdivia, Dubravka	108, 115, 130, 165
Swain, Matthew	092
Swiggett, Wanda	028

T

NAME	SESSION NUMBER
Tan, Edna	080
Tan, Yanyan	070, 161
Tang, Nai-En	082, 114
Tang, Xiuxiu	155
Taylor, Darius D.	096
Taylor, Melinda Ann	120
Templin, Jonathan	158
Tenison, Caitlin	070, 075
Tenorio, Alvin Daiz	034
Terada, Takeshi	154
Terao, Takahiro	026
Thompson, Kathryn Nicole	091, 125, 131
Thompson, Phylicia	126
Thompson, William Jacob	121, 143, 153
Thorp, Anna	110
Thum, Yeow Meng	069
Thurlow, Martha	123
Tian, Chen	016
Tomkowicz, Joanna	022, 109, 114
Tong, Ye	142, 146
Topczewski, Anna	078
Toton, Sarah Linnea	010
Toutkoushian, Emily Karen	072
Trippe, Matt	023, 092
Tsai, Chia-Lin	082, 121
Tseng, Ching-Yu	106
Tu, Naidan	082
Twing, Jon S.	083, 160
Tyack, Lillian	065

U

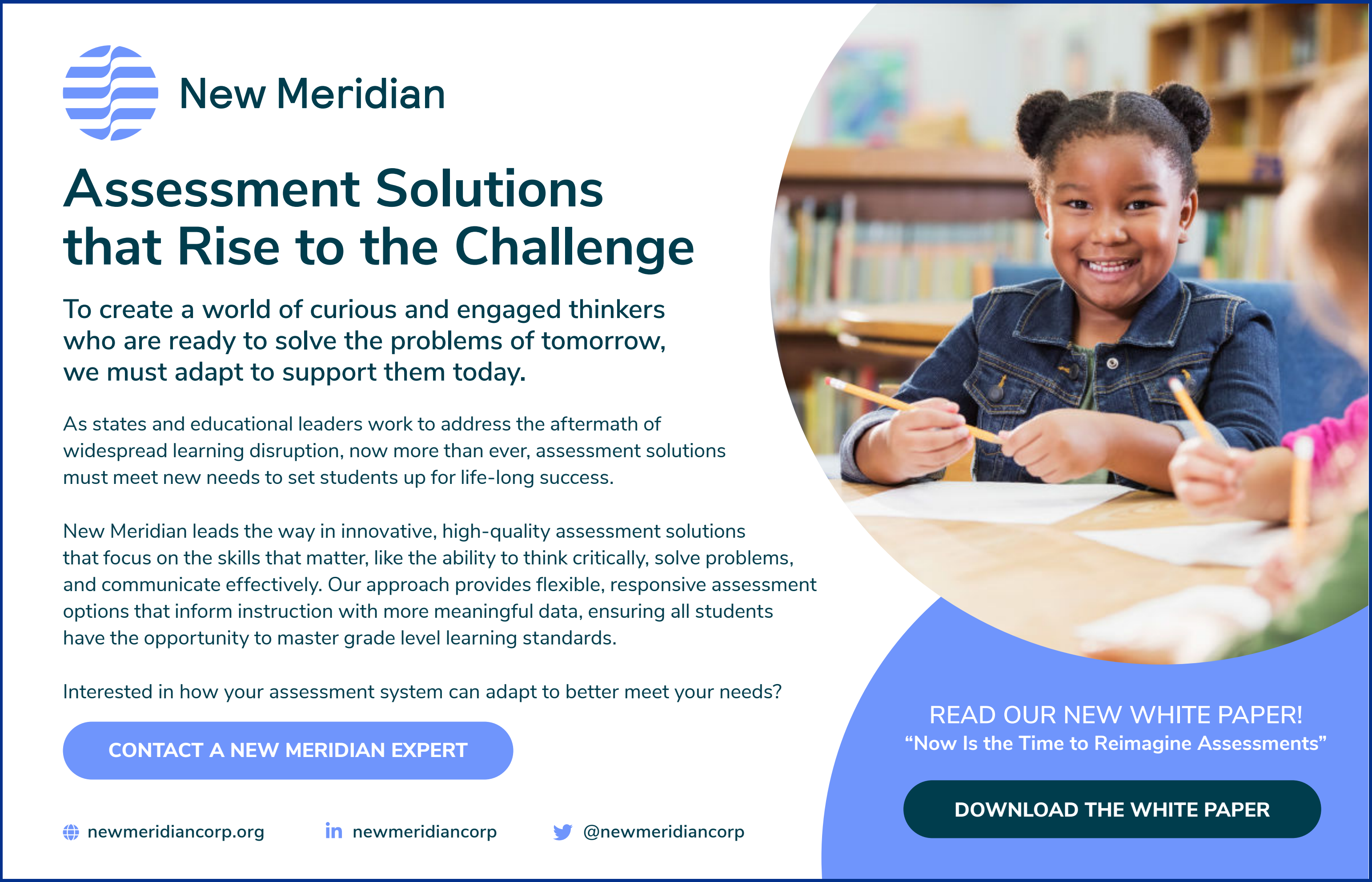
NAME	SESSION NUMBER
Ulitzsch, Esther	064, 077, 129, 165
Ulitzsch, Vincent	129
Umbricht, Mark	075

V

NAME	SESSION NUMBER
Vafis, Gregory	073
Valdivia Medinaceli, Montserrat B	108, 130
VanIwaarden, Adam	040, 043, 061, 132
Van Moere, Alistair	162
Varas, Leonor	112
Veldkamp, Bernard	057
Verkuilen, Jay	091
Vernon, Annette	151
Vispoel, Walter	026, 111
Vo, Thao	106, 114, 139
Vo, Yen	151
von Davier, Alina A.	053
Von Davier, Matthias	065
Vorenkamp, Ellen	102

W

NAME	SESSION NUMBER
Walker, Marcus	027, 047
Walker, Michael E.	020, 120, 150
Walkowiak, Temple A.	110
Wallace, Glen	023
Walters, Kristy	102
Wan, Ping	114
Wan, Siyu	085
Wang, Aijun	130
Wang, Bowen	162
Wang, Chun	063, 086, 098, 135, 153, 166
Wang, Chunxin	017
Wang, Dongwei	149
Wang, Hongling	149
Wang, Huan	022
Wang, Jiaxi	032, 039, 042
Wang, Kuo	164
Wang, Pujue	062
Wang, Selena	152
Wang, Shichao	106
Wang, Shiyu	011, 070, 161
Wang, Shudong	020
Wang, Songtao	122
Wang, Tianyou	118
Wang, Ting	147
Wang, Wei	150
Wang, Weimeng	031, 063
Wang, Wenhao	121
Wang, Xi	113, 144
Wang, Yu	039, 042
Way, Walter	076
Weber, Anjali	099
Weeks, Jonathan	064, 067, 111
Wei, Hsin-Ro	114
Wei, Jing	162



National Council on Measurement in Education is very grateful to the following organizations for their generous financial support of our 2022 Annual Conference

GOLD

ACT®

duolingo english test

ETS®

New Meridian

NBME®

Pearson

SILVER

ACS VENTURES
BRIDGING THEORY & PRACTICE

Alpine Testing Solutions

CA Cambium Assessment

CollegeBoard

edCount LLC®
because all students count

flexMIRT®
Vector Psychometric Group

University of Minnesota
Driven to Discover®

FRIEND

HARVARD
VE RI TAS
GRADUATE SCHOOL OF EDUCATION

HumRRO®
HUMAN RESOURCES RESEARCH ORGANIZATION

UMassAmherst
College of Education
Center for Educational Assessment

NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION
19 Mantua Road | Mount Royal, NJ 08061 | 856.284.3700
www.ncme.org