ANNUAL MEETING

EDUCATIONAL MEASUREMENT: IN SERVICE OF *Society*

DENVER 2025 ANNUAL MEETING ncme

APRIL 23-26, 2025

WELCOME FROM THE PROGRAM CHAIRS

Welcome to the 2025 NCME Annual Meeting! We are so excited to have you join us in Denver.

The theme of the 2025 NCME Annual Meeting is Educational Measurement: *In Service of Society*. The theme emphasizes the "why" of our work in educational measurement and echoes the last three words of the NCME mission: "to advance theory and applications of educational measurement to benefit society." We improve and innovate theory and practice not simply to advance the field but to serve society, such as with research that propels changes in educational policy and practice to improve learning and well-being. The extent to which our research benefits society depends on many factors, including the appropriateness of our methods, the richness of our partnerships, the nuance of our communications, and the quality of our data visualizations. Rigorous research coupled with outreach and effective communication are key ingredients in fulfilling our mission.

We invite you to attend the training and in-person presentations and engage in conversations with others on how measurement theory and practice may be reconceptualized to better serve society. Our hope for the 2025 Annual Meeting is to give attendees several ways to think about the theme and have further conversations on related measurement topics and research. There are a few events of the Annual Meeting that we wanted to highlight.

NCME 2025 includes over 20 different training sessions on Wednesday April 23. There will be half-day (4-hour) and mini (2-hour) sessions.

The first presentation day, **Thursday April 24**, will be a full day of great sessions organized in numerous formats: Coordinated Paper Sessions, Organized Discussions, and Individual Paper Sessions. There will also be three types of eBoard sessions, for Individual Paper, Graduate Student Paper, and Innovation Demonstration presentations. This day's program will end with two Joint NCME/AERA-Divisions D receptions: the Welcome Reception, followed by the Graduate Student Reception.

On the second presentation day, **Friday April 25**, we can look forward to the NCME Business Meeting and Presidential Address, starting at 7:30 AM followed by the NCME Career Award Address by Terry Ackerman at 9:45 AM.

The final presentation day, **Saturday April 26**, will start with the NCME Fitness 5k/2.5k Walk/Run, and end with the Informal Presidential Closing Reception, the location of which will be announced during the Business Meeting and Presidential Address. We hope you are able to join in many of these activities.

To give some more detail on the program, below are just a small number of selected sessions that attendees might be interested in each day.

Thursday April 24 Conference Sessions:

- Implications of Culturally Responsive Assessment for Large-Scale Assessment Practices (NCME Edited Volume: 11:30 AM)
- Testing Standards Revision: Focus on Foundations, Operations, & Psychometrics (1:30 PM)
- Joint NCME/AERA Division D Welcome Reception (6:30 PM)
- Joint NCME/AERA Division D Graduate Student Reception (8:30 PM)

Friday April 25 Conference Sessions:

- NCME Business Meeting & Presidential Address (7:30 AM)
- NCME Career Award (9:45 AM)
- NCME Task Force on the Role of Educational Measurement and Civil Rights (11:30 AM)
- Setting the Context for the Fifth Edition of Educational Measurement (1:30 PM)
- Supporting Social and Emotional Learning in Schools and Out-of-school-time Programs: Voices from the Field (1:30 PM)

WELCOME FROM THE PROGRAM CHAIRS

Saturday April 26 Conference Sessions:

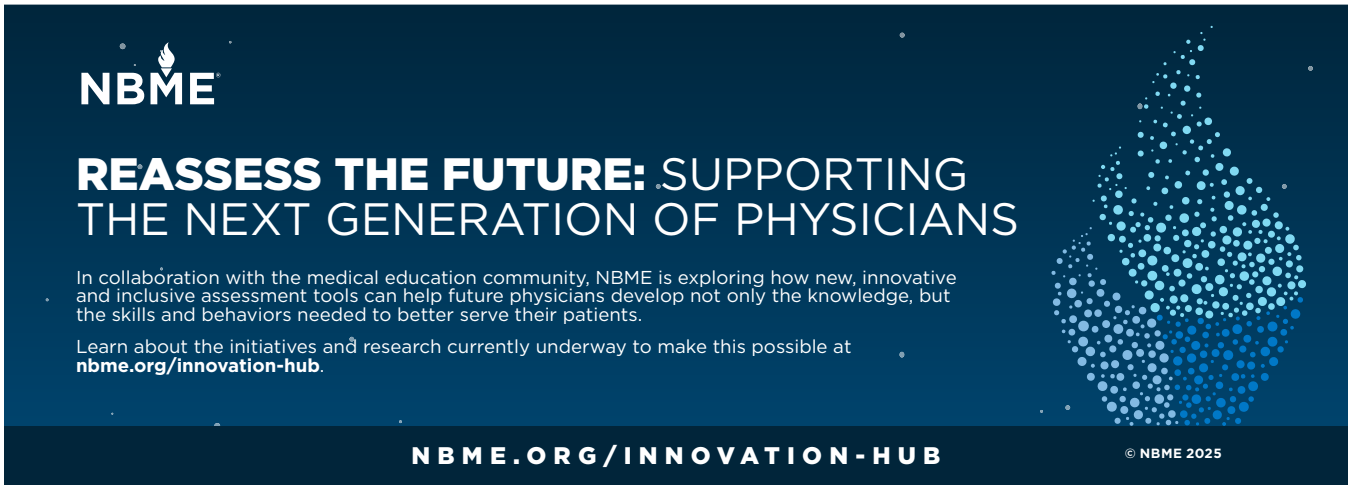
- NCME Fitness 5k/2.5k Walk/Run (6:00 AM)
- Testing Standards Revision: Focus on Applications of Measurement (8:00 AM)
- The Future of K12 Assessment & Accountability: Can Recent Research Inform Policy Changes? (11:30 AM)
- The Fifth Edition of Educational Measurement: Specific Applications of Assessment (1:30 PM)
- Informal Presidential Closing Reception (4:30 PM)

Electronic Board Sessions across Days:

- Graduate Student eBoard Sessions 1-5
- Innovation Demonstration Sessions 1 and 2
- Electronic Board Sessions 1-6 and Electronic Board Session: AI

We are so thankful to all who have contributed to this engaging program, including submission authors and all volunteers. We are appreciative of the reviewers for providing helpful feedback as well as colleagues who volunteered as discussants and chairs. We want to thank Qiwei (Britt) He and Matthew Madison, the Training and Professional Development Committee Co-Chairs, as well as Nicolas Mireles and Catherina Villafuerte, the Graduate Students Issues Committee Co-Chairs, for all of their work on the program. We are also very thankful to the previous NCME program chairs (Jonathan Weeks and Adam Wyse) for their help as well as feedback and discussion with next year's program chairs (Stefanie Wind, Pamela Kaliski) and president (Amy Hendrickson). Finally, we want to thank the NCME President Andrew Ho and NCME Executive Director Rich Patz for their help, support, and encouragement.

Katherine Castellano and Scott Monroe
2025 NCME Annual Conference Co-Chairs



Assessments that inform. Data that empowers.

The Pearson Assessment for Learning Suite is a category of K-12 assessments that support educators, caregivers, and students in improving their educational outcomes. Within the suite, our products include state standards-based measures for interim and formative purposes, skill-based screeners and progress monitoring tools to support MTSS/RTI programs, and additional screeners for early literacy, dyslexia, and behavior. Together, as part of a system of assessments, they form powerful insights into the needs of students.

Visit **PearsonAssessments.com/Learning-Suite** to learn more.

800-627-7271 | PearsonAssessments.com

Copyright © 2025 Pearson Education, Inc. or its affiliates. All rights reserved. Pearson is a trademark, in the US and/or other countries, of Pearson plc. C12559 SR 03/25

Pearson

ARTIFICIAL INTELLIGENCE IN MEASUREMENT & EDUCATION CONFERENCE

October 27 - 29, 2025 | Pittsburgh, Pennsylvania

SAVE THE DATE

We are thrilled to announce an NCME Special Conference called the ***Artificial Intelligence in Measurement and Education Conference (AIME–Con)***. It will be held in-person in **Pittsburgh, PA** on **October 27–29, 2025**, at the Wyndham Grand Pittsburgh Downtown. The theme is "**Innovation and Evidence: Shaping the Future of AI in Educational Measurement**".

This conference will gather an interdisciplinary group of experts from education, artificial intelligence, measurement, natural language processing, assessment, learning analytics, and technological development to foster learning across all groups.

The AIME–Con will include training sessions on October 27, and keynote, poster, and paper presentations on October 28 and 29. We welcome individuals to sign up to facilitate training sessions and review proposals. **Proposal submissions for paper and training sessions will open in early spring and are due June 15, 2025.** Notifications of acceptances will be sent in July when conference registration will open.

For more information, please see the conference website. We are excited to see you in Pittsburgh in October!

TABLE OF CONTENTS

General Meeting Information 07

Floor Plans 08

NCME Leadership 10

Proposal Reviewers 13

Schedule at a Glance 18

TRAINING SESSIONS

Wednesday, April 23 26

ANNUAL MEETING PROGRAM

Thursday, April 24 38

Friday, April 25 94

Saturday, April 26 138

Our Belief Statement

Every individual brings unique experiences, skill sets, and perspectives that work to advance our purpose: continuously improving the quality, fairness, and accessibility of education for all students.

Our Services

- Assessment Design, Development, and Evaluation
- Instructional Systems and Capacity Building
- Policy Analysis and Technical Assistance

edCount® LLC
because all students count

www.edCount.com
(202) 895-1502 | info@edCount.com

GENERAL MEETING INFORMATION

Welcome to the 2025 NCME Annual Meeting in Denver, CO!

NCME/AERA REGISTRATION & INFORMATION DESK

NCME check-in and registration is located on the Lower Level 2 foyer at the Hilton Denver City Center. Stop by the registration desk to pick up your conference materials including your name badge.

NCME Registration will be open the following hours:

- Wednesday, April 23, 7:30 am - 5:00 pm
- Thursday, April 24, 7:30 am - 5:00 pm
- Friday, April 25, 7:30 am - 5:00 pm
- Saturday, April 26, 7:30 am - 12:00 pm

NCME Gives Back

Every year, the NCME President in consultation with the Board of Directors selects a local charity to support. This year's selected charity is the Denver Scholarship Foundation. The Foundation uses evidence-based support systems and scholarships in partnership with the Denver Public Schools to increase college access and completion. Last year, the Foundation served 12,000 DPS students and nearly 2,200 selected scholars. You can read more about their work in their annual report. I have donated myself, and I encourage NCME members to support this foundation or similar charities as they are inclined and able.

Andrew Ho, President

National Council on Measurement in Education

DOWNLOAD THE OFFICIAL NCME EVENTS APP

- Download the eConference.io app from the App Store or Google Play.
- Open the app and enter NCME 2025 as the conference code.
- Once you are in the app, "Click to Log In" at the top of the screen and enter the log-in credentials from the Know Before You Go email to access.

X

Share your experience at the NCME Annual Meeting by using #NCME2025

FLOOR PLANS

LOWER LEVEL 1

LOWER LEVEL 2

FLOOR MAPS KEY

- Meeting/Conference
- Amenities
- Public Space
- Private

The AAMC Strengthens the World's Most Advanced Medical Care

AAMC
Tomorrow's Doctors, Tomorrow's Cures®

Representing institutions that deliver the world's most advanced medical care, the AAMC provides programs and services that support the entire spectrum of education, research, and health care. We believe in medical education that prepares physicians and scientists to meet the nation's evolving health needs, support a diverse and culturally competent health care workforce, lead medical advancements that prevent disease and alleviate suffering, and more.

Learn more at aamc.org.

24-061 (03/24)

Association of
American Medical Colleges

ACS VENTURES
BRIDGING THEORY & PRACTICE

www.acsventures.com
acs-ventures-llc
inquiries@acsventures.com

Professional Certification • Professional Licensure • Education

ACS was established to address a need in the assessment community for design, evaluation, operational support, and quality assurance. Our staff members consult directly with state, regional, national, and international education and credentialing testing programs. ACS has designed and led numerous test development and validation activities, conducted alignment and standard setting studies for a wide variety of clients, and has provided actionable and appropriate guidance and feedback to program leaders on various testing related issues.

Design

Program Design & Redesign
Assessment Design & Redesign
Research Design

Operational Support

Standard Setting
Alignment
Psychometric Analysis
Technical Documentation

Quality Assurance

Psychometric Oversight
Program Review and Audit
Program Evaluation
Technical Advisory Committees

NCME LEADERSHIP

NCME OFFICERS

President	**Andrew Ho**, *Harvard University*
President-Elect	**Amy Hendrickson**, *The College Board*
Past President	**Michael Walker**, *HumRRO*

NCME DIRECTORS

Leslie Keng
National Board of Medical Examiners

Li Cai
UCLA- CRESST

Brad McMillen
Wake County Public Schools

Zachary Warner
New York State Education Department

Susan Davis-Becker
ACS Ventures

Rochelle Michel
Smarter Balanced

NCME EXECUTIVE DIRECTOR

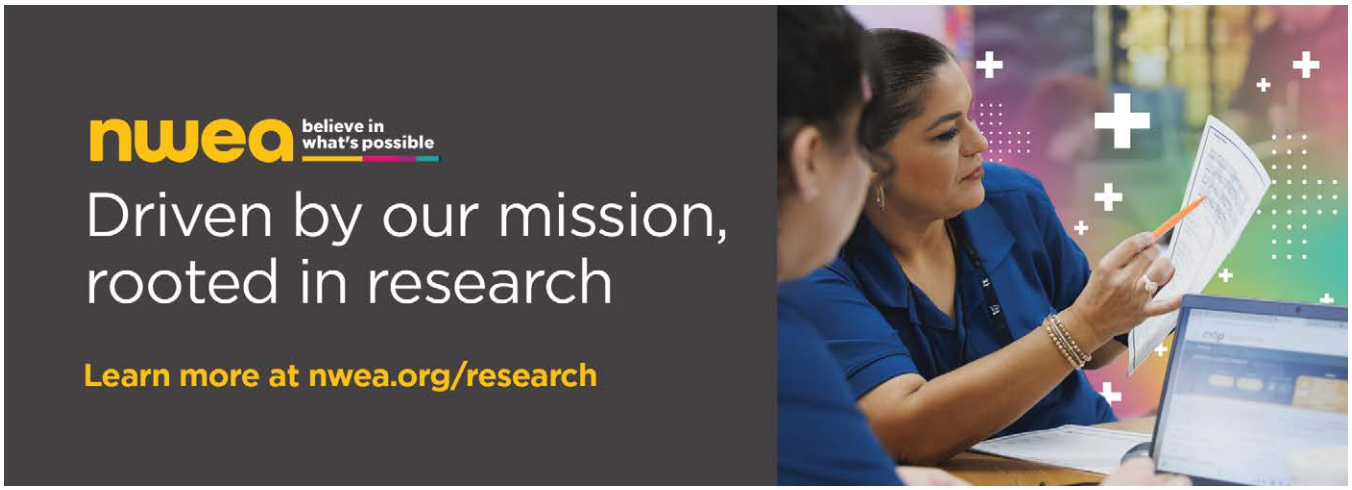
Richard J. Patz

EDITORS

Journal of Educational Measurement	**Chun Wang** *University of Washington*
Chinese/English Journal of Educational Measurement and Evaluation	**Yi Zheng** *Arizona State University* **Hua-Hua Chang** *Purdue University* **Tao Xin** *Beijing Normal University*
Educational Measurement: Issues and Practice	**Heather Buzick** *ETS*
ITEMS Editor	**Stella Kim** *University of North Carolina at Charlotte*
NCME Book Series Editor	**Kadriye Ercikan** *ETS*
News@NCME Editor	**Cheng Hua** *University of Montevallo*

2025 ANNUAL MEETING CHAIRS

NCME Website Editor	**Nathan Dadey** *Center for Assessment*
Annual Meeting Program Chairs	**Katherine Furgol Castellano** *ETS* **Scott Monroe** *University of Massachusetts Amherst*
Training and Professional Development Committee Chairs	**Qiwei Britt He** *Georgetown University* **Matthew J. Madison** *The University of Georgia*
Graduate Student Sessions Chairs	**Catherina Villafuerte** *University of Connecticut* **Nicolas Mireles** *James Madison University*
Fitness Run/Walk Director	**Katherine Furgol Castellano** *ETS*



The Problem: A third of college students end up spending time and money on remedial math courses that don't count toward their degrees because they lack sufficient math skills.

The Solution: Precalculus is one of the most powerful math courses in American high schools — taking it increases a student's likelihood of completing a bachelor's degree by 155%.

The Power of AP® Precalculus

AP® Precalculus is a new course designed to prepare more students for the math they'll encounter in college, and it can be offered in place of your current precalculus or trigonometry course. AP Precalculus students can earn college credit, advanced placement, or both. AP Precalculus students have more time with their teacher to better develop content knowledge and skills. They get 140 hours of time with their AP teacher versus about 48 hours in a traditional college course and in a smaller, familiar setting.

Learn more at **cb.org/ap-precalculus**

© 2023 College Board.

PROPOSAL REVIEWERS

Ketan
Doris Abroampah
Merve Sahin Kursad
Patrìcia Martinkov·
Christiana Akande
Tony Albano
Ernest Amoateng
Lily An
Benjamin Andrews
Sergio Araneda
Nana Amma Asamoah
Elizabeth Ayers-Wright
Paul Bailey
David Bamat
Tanushree Banerjee
Erin Banjanovic
Yu Bao
Jonathan Beard
Lee Becker
Ahmed Bediwy
Anton Beguin
Magdalen Beiting-Parrish
Yoav Bergner
Tanesia Beverley
Ummugul Bezirhan
Jared Block
Obed Boateng
Daniel Bolt
Wes Bonifay
Katrina Borowiec
Aimee Boyd
Richard Brown
Chad Buckendahl
Michael Bunch
Kayla Burt
Heather Buzick
Sandra Liliana Camargo Salamanca
Ian Campbell
Luciana Cancado
Maritza Casas
Roti Chakraborty
Shu-Ren Chang
Jianshen Chen
Qipeng Chen
Yiling Cheng
Yi-Chen Chiang
Hye-Jeong Choi
Jinah Choi
Ikkyu Choi
Catherine Close
Kimberly Colvin
Steven Culpepper
Laurie Davis
Susan Davis-Becker
Teresa (Tess) Dawber
Lawrence DeCarlo
Onur Demirkaya
Jiayi Deng
Benjamin Domingue
John Donoghue
Bryan Drost
Aybuke Durmus
Carol Eckerly
Ejike Edeh
Daniel Edi
Kelly Edwards
Farshad Effatpanah Hesari
Kerry Englert
Fen Fan
Meng Fan
Ziyu Fan
Michael Fauss
Richard Feinberg
Tianying Feng
Zechu Feng
Anthony Fina
W. Holmes Finch
John Fremer
Brian French
Andreas Frey
Yanyan Fu
Melissa Gallagher
Brian Gane
Yuan Ge
Joshua Goodman
Rumeysa Gorgulu
Kylie Gorney
Edith Graf
Nikole Gregg
Yage Guo
Ahmet Guven
Sarah Hagge
Suhwa Han
Mostafa Hanafy
Mark Hansen
Surina He
Tracey Hembry
Amy Hendrickson
Igor Himelfarb
Kari Hodge
Alexander Hoffman
Hyeri Hong
Youmin Hong
Yue Huang
Yingshi Huang
Mingya Huang
Corinne Huggins-Manley
Aria Immanuel
Janine Jackson Haynes
Miriam James
Minjeong Jeon
Paul Jewsbury
Jeongmin Ji
Xuejun Ji
Yanming Jiang
Kuan-Yu Jin
Matthew Johnson
Joanne Joo
Unhee Ju
Juyoung Jung
Nilufer Kahraman
Priya Kanaan
Hyeon-Ah Kang
Klint Kanopka
Cayce Kantor
Shalini Kapoor
Hacer Karamese
Tzur Karelitz
Russell Keglovits
Lisa Keller
Emily Kerzabi
Marisol Kevelson
Stella Kim
Yun-Kyung Kim
Young Yee Kim
Seongeun Kim
Han Yi Kim
Nana Kim
Youngwon Kim
Alexandra Kislyonkova
Jennifer Kobrin
Charalambos Kollias
Andrew Krist
Marc Kroopnick
Kevin Krost
Ferdinand Kumolalo
Tzu-Chun Kuo
Joseph Kush
Alexander Kwako

PROPOSAL REVIEWERS

Jungwon Kyung
Suzanne Lane
Meredith Langi
Anh Thu Le
Soo Lee
Haeju Lee
Hyeryung Lee
Hyunjung Lee
Mina Lee
Won-Chan Lee
Blair Lehman
Adam Lekwa
Dihao Leng
Brian Leventhal
Daniel Lewis
Jennifer Lewis
Dongmei Li
Jing Li
Min Liang
Xiangyi Liao
Yuan-Ling Liaw
Youn Seon Lim
Ye Lin
Yuming Liu
Huan Liu
Liu Liu
Xiang Liu
Yunting Liu
Chunyan Liu
Susan Lottridge
Quintin Love
Yikai Lu
Ru Lu
Ying Lu
Benjamin Kweku Lugu
Jinwen Luo
Wanjing (Anya) Ma
Wenchao Ma
Hotaka Maeda
Jaime Malatesta
Kaiwen Man
Preeti Manchanda
Constanza Mardones-Segovia
Scott Marion
Maura Maxfield
Daniel McCaffrey
Amirreza Mehrabi
Roghayeh Mehrazmay
Fernando Mena
Huijuan Meng
Alejandra Miranda
Nicolas Mireles
Robert Mislevy
Sebastian Moncaleano
Kristin Morrison
Marek Muszyński
Aaron Myers
Kelsey Nason
Sebastian Nastuta
Milly Njeri
Kate Nolan
Steven Nydick
Alexis Oakley
Teresa Ober
Hyeonjoo Oh
Comfort Omonkhodion
Christopher Ormerod
Daniel Oyeniran
Fatih Ozkan
Seohong Pak
Jose Palma
Qianqian Pan
Seohee Park
Hyemin Park
Thanos Patelis
Chris Patterson
Richard Patz
Marianne Perie
Wallace Pinto
Sonya Powers
Victoria Quirk
Onur Ramazan
He Ren
Michael Rodriguez
Oxana Rosca
Louis Roussos
Andrew Runge
Christopher Runyon
Andre Rupp
Leslie Rutkowski
Godwin Sabboh
Edgar Sanchez
Merve Sarac
Paulius Satkus
Edynn Sato
Ayfer Sayin
Madeline Schellman
Bob Schwartz
Fajer Shamsaldeen
Yelisey Shapovalov
Benjamin Shear
Hyejin Shim
Julian Siebert
Sandip Sinharay
Gozde Sirganci
William Skorupski
Jessalyn Smith
James Soland
Dan Song
Parth Soni
Jayashri Srinivasan
Dorota Staniewska
Jeffrey Steedle
Kelley Stethen
Sanford Student
Yu-Lan Su
Yon Soo Suh
Youmi Suk
Jeneve Swaby
Bin Tan
Victoria Tanaka
Cheng Tang
Lissette Tolentino
David Torres Irribarra
Emily Toutkoushian
Colt Turner
Jon Twing
Secil Ugurlu
Monika Ulrich
Emmanuel Umeh
Stephanie Underhill
Montserrat Valdivia Medinaceli
Dustin Van Orman
Peter van Rijn
Sophia Velez
Annette Vernon
Yen Vo
Ting Wang
Aijun Wang
Shichao Wang
Bowen Wang
Xi Wang
Min Wang
Zifan Wang
Tao Wang
Shiyu Wang
Jonathan Weeks
Craig Wells
Paul Westrick
Jordan Wheeler

PROPOSAL REVIEWERS

John Whitmer
Andrew Wiley
Caroline Wiley
Benjamin Williamson
Joshua Wilson
Stefanie Wind
Scott Wood
Tong Wu
Yi-Fang Wu
Yi-Chen Wu
Adam Wyse
Yangmeng Xu
Xiaochen Xu
Rujun Xu
Mingfeng Xue
Guanao Yan
Ji Seung Yang
Lihua Yao
Sinan Yavuz
Shu Jing Yen
Hanwook Yoo
Haoyang Yu
Diego Zapata-Rivera
Paul Zavitkovsky
Biao Zeng
April Zenisky
Peida Zhan
Yuxiao Zhang
Xiuyuan Zhang
Yifan Zhang
Fan Zhang
Xinchang Zhou
Xuechun Zhou
Sizheng Zhu
Rongchun Zhu
Selay Zor

TRAINING SESSION REVIEWERS

Yu Bao
Bozhidar Bashkov
Tabitha Bellamy
W. Holmes Finch
Fernanda Gandara
Suhwa Han
Jiangang Hao
Sijia Huang
Hong Jiao
Hyeon-Ah Kang
Huan Kuang
Andrew Lan
Dongmei Li
Liyuan Liu
Ray Reichenberg
He Ren
Frank Rijmen
Javier Su·rez-¡lvarez
Youmi Suk
Xueying Tang
Mark Wilson
Yi-Fang Wu
Ye Yuan
Yi Zheng

GRADUATE STUDENT POSTER COMPETITION JUDGES

Sergio Araneda-Galarce
Kayla Burt
Jennifer Dunn
Brian French
Brian Leventhal
Nicolas Mireles
Caroline Prendergast
Julian Siebert
Lissette Tolentino
Anne Traynor
Montserrat Valdivia Medinaceli
Catherina Villafuerte
Thao Vo
Michael Walker
Jiawei Xiong

www.nciea.org

31 MOUNT VERNON STREET • DOVER, NH 03820 • 603.516.7900

Center for Assessment
The National Center for the Improvement of Educational Assessment, Inc.

duolingo english test

Innovation in assessment

The Duolingo English Test leverages the latest assessment science and human-in-the-loop AI to empower anyone to test where and when they're at their best.

Digital-first
The Duolingo English Test leverages AI to personalize itself in real-time to every test taker, honing in on their true proficiency more quickly, precisely, and securely than traditional fixed-form tests.

Accessible administration
The Duolingo English Test is designed to radically improve English language proficiency assessment for test takers and score recipients alike, by providing a testing experience that is accessible while remaining accurate and secure.

Secure certification
Unprecedented security technology and a 1:1 candidate-to-proctor ratio ensure that the Duolingo English Test is extremely secure. Every proctor has access to AI tools that monitor dozens of categories of biometrics and behavioral data.

englishtest.duolingo.com/research

Explore our research

Keep up to date on the latest DET news

The future of *i-Ready Assessment* is invisible.

- Voice technology is coming to *i-Ready Literacy Tasks*.
- Built to hear students' voices of all accents and dialects
- Creating the best possible solution by collaboratively learning with teachers in the classroom

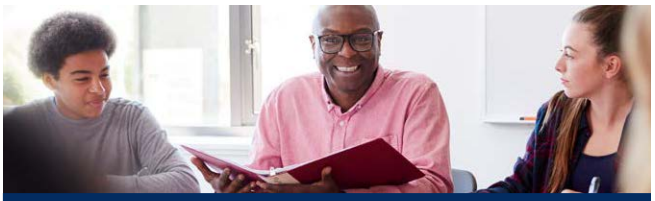
SoapBoxLabs
A Curriculum Associates Company

Learn about our vision for the future with *i-Ready Literacy Tasks*.

DATA RECOGNITION DRC CORPORATION

DRC IS HONORED TO SPONSOR THE 2025 NCME CONFERENCE

www.datarecognitioncorp.com



Providing insights on education and workplace success through innovative and collaborative research

ACT Research leads the field by producing high-quality scientific evidence in support of solutions for education and workforce readiness.

Our research teams provide policymakers, educators, parents, and learners with research-based insights to inform their decision-making and deliver educators and workforce development professionals with tools and serviced needed for education and career navigation.

Learn more at act.org/research

ACT

SCHEDULE AT-A-GLANCE

(SESSIONS MARKED WITH AN * HAVE AN ADDITIONAL COST)

Wednesday, April 23

Begin Time	End Time	Room	Session Title
8:00 AM	12:00 PM	Penrose 1	A Brief Introduction to Big Data Techniques Using International Large-Scale Assessment Data*
8:00 AM	12:00 PM	Colorado AB	An Overview of Operational Psychometric Work in Real World*
8:00 AM	12:00 PM	Denver 4	Applications of Bayesian Networks in the Age of AI*
8:00 AM	12:00 PM	Penrose 2	Applying Data Mining Methods to Detect Test Fraud*
8:00 AM	12:00 PM	Colorado CD	Best Practices for Artificial-Intelligence Scoring*
8:00 AM	10:00 AM	Denver 3	Build Your Personal Website With Quarto*
8:00 AM	12:00 PM	Colorado IJ	Comprehensive Statistical Model Evaluation*
8:00 AM	12:00 PM	Mattie Silks	Data Visualizations and Dashboards to Inform Assessment Development Using R*
8:00 AM	10:00 AM	Denver 1-2	Formative Assessments or the Formative Assessment Process – The -s Makes the Difference*
8:00 AM	12:00 PM	Denver 5-6	VEMIRT: An R Package for Conducting High-Dimensional IRT Analysis*
10:30 AM	12:30 PM	Denver 3	Funding Opportunities in Measurement from Institute of Education Sciences*
10:30 AM	12:30 PM	Denver 1-2	Harder than Rocket Science? Explaining Assessment to Everyday Audiences*
12:40 PM	2:10 PM	Bluebird Ballroom 3G	Using NAEP Data to Investigate Student Differences in Education (Joint NCME/AERA Session - Convention Center)
1:00 PM	5:00 PM	Denver 5-6	Building Monte Carlo Simulations in R*
1:00 PM	5:00 PM	Colorado IJ	Cloud Computing and Artificial Intelligence in Psychometric Applications*
1:00 PM	5:00 PM	Colorado AB	Cognitive Diagnosis Modeling: A General Framework Approach and Its Implementation in R*
1:00 PM	5:00 PM	Denver 3	Incorporating Process Data in Large-Scale Assessments: Methods and Case Studies*
1:00 PM	5:00 PM	Penrose 2	Introduction to AI-based Automated Item Generation and Automated Scoring*
1:00 PM	3:00 PM	Penrose 1	Modeling Item-Level Heterogeneous Treatment Effects: A Tutorial in R*
1:00 PM	5:00 PM	Denver 4	Opening the Black Box: Understanding Content Development for Large Scale Standardized Assessment*
1:00 PM	5:00 PM	Denver 1-2	Organizational Leadership for Measurement Experts*
1:00 PM	3:00 PM	Mattie Silks	Positionality in Action: Developing Tools to Recognize Bias Across Settings*
1:00 PM	5:00 PM	Colorado CD	Tools and Strategies for the Design and Evaluation of Interactive Dashboard Reports*

Wednesday, April 23

Begin Time	End Time	Room	Session Title
3:30 PM	5:30 PM	Mattie Silks	Culturally and Linguistically Responsive Assessment for Learning in STEM*
3:30 PM	5:30 PM	Penrose 1	PACER for Statistics and Psychometrics*

Thursday, April 24

Begin Time	End Time	Room	Session Title
8:00 AM	9:30 AM	Colorado IJ	Serving Diverse Student Groups Through Accommodation and Accessibility Support in Large-Scale Assessments
8:00 AM	9:30 AM	Penrose 1	Computer-supported Collaborative Social Emotion Learning and Assessment in Classroom
8:00 AM	9:30 AM	Denver 1-2	Generative AI in Education: Applied Ethics & Evaluating Model Performance (Joint NCME/AERA Session & SIGIMIE-Organized Session)
8:00 AM	9:30 AM	Denver 3	Data-intensive research in educational measurement using the Item Response Warehouse
8:00 AM	9:30 AM	Denver 4	Item Difficulty Modeling in Support of Test Development and Validation
8:00 AM	9:30 AM	Denver 5-6	Ensuring Fairness and Consistency in Assessment Practices
8:00 AM	9:30 AM	Colorado GH	Electronic Board Session 1
8:00 AM	9:30 AM	Colorado AB	DIF Applications
8:00 AM	9:30 AM	Mattie Silks	Student Growth Modeling
8:00 AM	9:30 AM	Colorado CD	New IRT Statistical Tools
8:00 AM	9:30 AM	Colorado EF	Connecting Performance Level Descriptors and Released Items to Improve Understanding and Communication (SIGIMIE-Organized Session)
8:00 AM	9:30 AM	Mattie Silks	Are Standardized Admissions Tests Fair? If Not, How Might We Make Them Fairer?
9:45 AM	11:15 AM	Denver 1-2	A Framework to Guide Development and Evaluation of AI-Based Formative Assessment Systems
9:45 AM	11:15 AM	Colorado AB	Enhancements to the ACT: A Summary of Research Evaluating Use of ACT Scores
9:45 AM	11:15 AM	Colorado EF	More than Words: Intersectionality in Education Research and Policy
9:45 AM	11:15 AM	Colorado IJ	Innovations in the Assessment of Civic Learning in K-12 Schools
9:45 AM	11:15 AM	Denver 5-6	Innovations in Using Large Language Models for Automated Scoring and Automated Writing Evaluation
9:45 AM	11:15 AM	Denver 3	IRT Scoring and Rating Methods
9:45 AM	11:15 AM	Colorado CD	Leveraging Through-Year/Interim Assessment Results for Instructional Use
9:45 AM	11:15 AM	Denver 4	Process Data Analyses to Understand Large-Scale Survey-Based Assessment Results

Thursday, April 24			
Begin Time	**End Time**	**Room**	**Session Title**
9:45 AM	11:15 AM	Colorado GH	Innovation Demonstration Session 1
9:45 AM	11:15 AM	Penrose 2	Is Unaccounted-for Multidimensionality Really That Problematic?
9:45 AM	11:15 AM	Penrose 1	What has, can, and should happen when the Standards are not followed? (NCME Committee Organized Session)
11:30 AM	1:00 PM	Penrose 2	Social-Emotional Learning Measurements in Service of ALL Students: Lessons Learned around COVID
11:30 AM	1:00 PM	Mattie Silks	Fuse Diverse Process Data in Educational Measurement
11:30 AM	1:00 PM	Denver 5-6	Exploring Methods to Expand Reporting on NAEP for Groups of Students
11:30 AM	1:00 PM	Denver 1-2	On the Nature and Rigor of Classroom Assessment
11:30 AM	1:00 PM	Colorado GH	Graduate Student eBoard Session 1
11:30 AM	1:00 PM	Denver 3	Improving Scoring for CAT
11:30 AM	1:00 PM	Colorado CD	Measurement Invariance and Differential Item Functioning
11:30 AM	1:00 PM	Denver 4	Rapid Guessing and Aberrant Responses
11:30 AM	1:00 PM	Colorado IJ	Implications of Culturally Responsive Assessment for Large-Scale Assessment Practices (NCME Edited Volume: Invited Session)
11:30 AM	1:00 PM	Penrose 1	Test Security in the AI Age: Trends, Challenges, and Visions (SIGIMIE-Organized Session)
11:30 AM	1:00 PM	Colorado EF	Honoring the Contributions of Steve Ferrara: Cognitive Complexity
1:30 PM	3:00 PM	Denver 1-2	Applications of Multistage Testing in the Age of Digital and AI Technology
1:30 PM	3:00 PM	Denver 3	Gaining New Insights Through Analysis of Multimodal, Multi-Source Data
1:30 PM	3:00 PM	Colorado AB	Automating Item Development with Artificial Intelligence: From Generation to Evaluation
1:30 PM	3:00 PM	Colorado GH	Graduate Student eBoard Session 2
1:30 PM	3:00 PM	Denver 5-6	Comparing Approaches for Using Large Language Models for Scoring and Specialized Assessments
1:30 PM	3:00 PM	Colorado IJ	Longitudinal/Multilevel Modeling
1:30 PM	3:00 PM	Penrose 2	Measuring Complex Skills
1:30 PM	3:00 PM	Colorado CD	Teacher Evaluation and School Accountability
1:30 PM	3:00 PM	Penrose 1	Barrier and Challenges for Transitioning Students from AA-AAAS to General Assessments
1:30 PM	3:00 PM	Mattie Silks	Exploring Effective, Equitable Classroom Assessment to Support Teacher Practice and Student Success (NCME Committee Organized Session)
1:30 PM	3:00 PM	Colorado EF	Testing Standards Revision: Focus on Foundations, Operations, & Psychometrics

Thursday, April 24			
Begin Time	**End Time**	**Room**	**Session Title**
3:15 PM	4:15 PM	Colorado EF	Progress and Pathways: Black Students, Scholars, and Sages in Measurement
3:15 PM	4:15 PM	Colorado IJ	Research on Remote Proctoring in Support of Valid Test Scores
3:15 PM	4:15 PM	Colorado AB	How Scoring Decisions Can Introduce Severe Bias into Growth Mixture Model Results
3:15 PM	4:15 PM	Penrose 2	Applications of Generative AI to Mathematics Education: Opportunities and Challenges
3:15 PM	4:15 PM	Colorado GH	Electronic Board Session 2
3:15 PM	4:15 PM	Denver 4	Using AI to Address Challenges in Scoring and Automated Feedback
3:15 PM	4:15 PM	Denver 1-2	Dimensionality/Factor Analysis
3:15 PM	4:15 PM	Denver 5-6	IRT Item Calibration
3:15 PM	4:15 PM	Colorado CD	Modeling Process Data
3:15 PM	4:15 PM	Denver 3	Score Reporting
3:35 PM	5:05 PM	Room 705	Toward Responsible Use of AI in Education Research and Development (Joint NCME/AERA Session - Convention Center)
4:30 PM	6:00 PM	Colorado AB	Generative-AI automatic item generation for large-scale language assessments
4:30 PM	6:00 PM	Penrose 2	Examinee Preknowledge: Understanding the Impact, Mitigating the Implications, and Detecting the Imprint
4:30 PM	6:00 PM	Colorado CD	Item Bank Maintenance in Three-Dimensional Science Assessment Programs
4:30 PM	6:00 PM	Denver 3	Using AI and Machine Learning to Improve Data on Students' Educational Progress in the US and Worldwide
4:30 PM	6:00 PM	Colorado GH	Electronic Board Session 3
4:30 PM	6:00 PM	Denver 5-6	Bayesian Inference for Latent Variable Modeling
4:30 PM	6:00 PM	Colorado IJ	DCM Q-Matrix Methods
4:30 PM	6:00 PM	Denver 1-2	Algorithms for Test Development
4:30 PM	6:00 PM	Colorado EF	Bridging Graduate Students and Practice: Internships in Educational Measurement (SIGIMIE-Organized Session)
4:30 PM	6:00 PM	Penrose 1	One Hundred Years of Criticisms Against Standardized Testing
6:30 PM	8:00 PM	Colorado Foyer	Joint NCME/AERA-Division D Welcome Reception
8:30 PM	10:00 PM		Joint NCME/AERA-Division D Graduate Student Reception

Friday, April 25

Begin Time	End Time	Room	Session Title
7:30 AM	9:30 AM	Colorado EF	NCME Business Meeting & Presidential Address
9:45 AM	11:15 AM	Colorado EF	NCME Career Award
9:45 AM	11:15 AM	Colorado AB	AI Innovative Applications in Automated Scoring
9:45 AM	11:15 AM	Denver 5-6	Evaluating the Psychometric Impacts of Cultural Representation in Item Contexts
9:45 AM	11:15 AM	Penrose 2	Revolutionizing Big Data with Advanced AI and Machine Learning (SIGIMIE-organized session)
9:45 AM	11:15 AM	Denver 3	Comparing Approaches to Reduce Vulnerabilities and Improve Performance of Large Language Models in Assessment
9:45 AM	11:15 AM	Colorado IJ	Computerized Adaptive Testing
9:45 AM	11:15 AM	Colorado CD	Through-Year/ Interim Assessment Designs and Analyses
9:45 AM	11:15 AM	Denver 1-2	Content and Internal Structure Validity Evidence
9:45 AM	11:15 AM	Colorado GH	Innovation Demonstration Session 2
9:45 AM	11:15 AM	Penrose 1	How Can We Balance Assessment Systems to Enable Teacher Collaborative Learning?
11:30 AM	1:00 PM	Colorado EF	NCME Task Force on the Role of Educational Measurement and Civil Rights (NCME Committee Organized Session)
11:30 AM	1:00 PM	Colorado CD	Dive into Item and Form Development Challenges in Large-Scale Assessment (SIGIMIE-Organized Session)
11:30 AM	1:00 PM	Colorado IJ	Approaches to Optimizing a Personalized Learning System
11:30 AM	1:00 PM	Colorado AB	Innovations in Automated Scoring Modeling
11:30 AM	1:00 PM	Penrose 1	Improving Score Reporting of Student Group Results to Better Support Educational Equity
11:30 AM	1:00 PM	Colorado GH	Graduate Student eBoard Session 3
11:30 AM	1:00 PM	Denver 1-2	New Models and Methods for DCM
11:30 AM	1:00 PM	Denver 4	DEI and Language Learning
11:30 AM	1:00 PM	Denver 5-6	Scaling and Linking with IRT
11:30 AM	1:00 PM	Penrose 2	Federal Policy Facilitators and Barriers for Performance Based Assessment Systems: National Perspectives on a Case Study
1:30 PM	3:00 PM	Denver 1-2	Computational Psychometrics in the Wild: Scalable Procedures for Maintaining Test Scores
1:30 PM	3:00 PM	Colorado CD	Competency-Based Education: Theory, Framework, and Practice
1:30 PM	3:00 PM	Denver 3	It's Time To Evaluate State Accountability Systems
1:30 PM	3:00 PM	Colorado AB	LLM respondents in Measurement
1:30 PM	3:00 PM	Denver 4	Integrating Multimodal Process Data in Digital Assessments
1:30 PM	3:00 PM	Colorado EF	Setting the Context for the Fifth Edition of Educational Measurement (Joint NCME/AERA Session)
1:30 PM	3:00 PM	Colorado GH	Graduate Student eBoard Session 4

Friday, April 25

Begin Time	End Time	Room	Session Title
1:30 PM	3:00 PM	Denver 5-6	Sampling Design and Imputation Method Advances for Large-Scale Survey-Based Assessments
1:30 PM	3:00 PM	Colorado IJ	Test Security: Methods and Considerations
1:30 PM	3:00 PM	Penrose 1	Re-humanizing Assessment: A Justice-Oriented Framework for Culturally Responsive Assessment Design (NCME Committee Organized session)
1:30 PM	3:00 PM	Penrose 2	Supporting Social and Emotional Learning in Schools and Out-of-school-time Programs: Voices from the Field
3:15 PM	4:15 PM	Denver 1-2	New Methods and Applications of DIF
3:15 PM	4:15 PM	Denver 3	The Unintended Consequences of Practical Scoring Decisions On Score Reliability and Interpretation
3:15 PM	4:15 PM	Denver 4	New Developments in Standard Setting
3:15 PM	4:15 PM	Colorado GH	Electronic Board Session 4
3:15 PM	4:15 PM	Penrose 2	Novel Quantitative Methods for Classroom Assessment
3:15 PM	4:15 PM	Denver 5-6	Examining Student Response Behavior in Large-Scale Survey-Based Assessments
3:15 PM	4:15 PM	Penrose 1	Predicting and Reporting Outcomes in Higher Education
3:15 PM	4:15 PM	Colorado AB	Using Process Data to Understand Student Response Behavior
3:15 PM	4:15 PM	Colorado CD	Evaluating Reliability, Precision, and Generalizability
3:15 PM	4:15 PM	Mattie Silks	Item Design and Item Difficulty
3:15 PM	4:15 PM	Colorado IJ	Test Security: Fradulent Behavior Detection Methods
4:30 PM	6:00 PM	Denver 3	Continuing Analyses of Frederic Lord's Randomly Parallel Testing
4:30 PM	6:00 PM	Colorado IJ	Recent Developments and Applications of Cognitive Diagnostic Models in Educational Measurement
4:30 PM	6:00 PM	Denver 1-2	Technology-based assessment of higher education students' critical online reasoning skills
4:30 PM	6:00 PM	Denver 5-6	Understanding Test-Taking (Dis)engagement: Predictors, Test Design, and Eye Movements in Low-Stakes Assessments
4:30 PM	6:00 PM	Colorado AB	Elevating Test Security in Validity Theory
4:30 PM	6:00 PM	Colorado CD	Using Large Language Models for Alignment and Progression Tasks
4:30 PM	6:00 PM	Denver 4	CAT Item Selection
4:30 PM	6:00 PM	Penrose 2	Content, Item, and Test Development
4:30 PM	6:00 PM	Colorado GH	Psychometric Overlord Auditions: Take 1
4:30 PM	6:00 PM	Penrose 1	Scaling, Linking, and Equating with Test Redesign (SIGIMIE-Organized Session)
4:30 PM	6:00 PM	Colorado EF	From Principles to Practice: Operationalizing Responsible AI in Educational Measurement
4:30 PM	6:00 PM	Gold Coin	Historically Marginalized Group Mixer (GSIC/CODIT)

Saturday, April 26			
Begin Time	**End Time**	**Room**	**Session Title**
6:00 AM	7:30 AM		NCME Fitness 5k/2.5k Walk/Run
8:00 AM	9:30 AM	Denver 4	States' Are Trying to Make Assessments More Instructionally Useful: How's It Working?
8:00 AM	9:30 AM	Denver 5-6	Recent Theoretical AI Developments in Educational Assessment
8:00 AM	9:30 AM	Colorado CD	Innovations in Latent Space Item Response Modeling: Diagnosing Hidden Dependencies for Fairness
8:00 AM	9:30 AM	Penrose 1	Fostering Equity In College Admissions: Test Reforms In Chile
8:00 AM	9:30 AM	Colorado AB	Investigating Validity of Learning Progressions in Computational Thinking (CoT)
8:00 AM	9:30 AM	Colorado GH	Electronic Board Session 5
8:00 AM	9:30 AM	Colorado IJ	Using Large Language Models within Item Development Workflows
8:00 AM	9:30 AM	Denver 3	Enhancing and Extending CAT
8:00 AM	9:30 AM	Denver 1-2	Modern Types of DIF
8:00 AM	9:30 AM	Penrose 2	Socioculturally Responsive Assessment and Measurement
8:00 AM	9:30 AM	Colorado EF	Testing Standards Revision: Focus on Applications of Measurement (Joint NCME/AERA Session)
9:45 AM	11:15 AM	Denver 1-2	Leveraging Large-Language Models for Operational Large-Scale Assessments: Opportunities and Challenges
9:45 AM	11:15 AM	Denver 5-6	Can Engagement and Measurement Precision be Enhanced in NAEP Mathematics through MST?
9:45 AM	11:15 AM	Colorado CD	Beyond Retrofitting: Flexibility and Comparability by Design for Equitable Assessment
9:45 AM	11:15 AM	Colorado AB	Pushing the Utility of Assessments in Competency-Based Education
9:45 AM	11:15 AM	Colorado IJ	Cutscore Estimation Via Two-Stage Analysis: Implications for Psychometric Research and Practice
9:45 AM	11:15 AM	Colorado GH	Electronic Board Session 6
9:45 AM	11:15 AM	Denver 4	Causal Models in Educational Measurement
9:45 AM	11:15 AM	Penrose 2	Innovations in Test Assembly
9:45 AM	11:15 AM	Denver 3	Generalizability Theory and Structural Equation Modeling
9:45 AM	11:15 AM	Penrose 1	Bridging Theory and Practice: Fostering Innovation in Diagnostic Measurement to Enhance Education (SIGIMIE-Organized Session)
9:45 AM	11:15 AM	Colorado EF	Psychometrically-driven Product Development in the Age of AI
11:30 AM	1:00 PM	Colorado CD	Socioculturally Responsive Assessment: Implications for Theory, Measurement, and Systems-Level Policy
11:30 AM	1:00 PM	Denver 3	Understanding AI and Human Ratings: Complementary Roles of AI and Human Raters
11:30 AM	1:00 PM	Denver 4	Using NAEP Data to Quantify and Address Educational Inequities

Saturday, April 26			
Begin Time	**End Time**	**Room**	**Session Title**
11:30 AM	1:00 PM	Penrose 2	Research for Practical Issues and Solutions in Computerized Multistage Testing (Routledge Book)
11:30 AM	1:00 PM	Colorado AB	Dynamic Measurement of Student Proficiency for Continuous Assessment
11:30 AM	1:00 PM	Colorado GH	Graduate Student eBoard Session 5
11:30 AM	1:00 PM	Denver 1-2	Expanding Automated Scoring Validity & Generalizability using AI
11:30 AM	1:00 PM	Denver 5-6	DCM for Longitudinal Data
11:30 AM	1:00 PM	Colorado IJ	Innovations in Technology-Based Assessment
11:30 AM	1:00 PM	Colorado EF	The Future of K12 Assessment & Accountability: Can Recent Research inform policy changes?
11:30 AM	1:00 PM	Penrose 1	The Usefulness of Kane's "Validity Argument" in Modern Validity Theory
1:30 PM	3:00 PM	Colorado CD	Advancements in Test Construction and Evaluation of Multistage Tests
1:30 PM	3:00 PM	Colorado AB	Building Just Assessments: A Culturally and Community-Grounded Perspective
1:30 PM	3:00 PM	Denver 4	Early Literacy Assessment: Developing, Validating, and Applying Measures that Support Student Learning
1:30 PM	3:00 PM	Colorado EF	The Fifth Edition of Educational Measurement: Specific Applications of Assessment (Joint NCME/AERA Session)
1:30 PM	3:00 PM	Colorado GH	Electronic Board Session: AI
1:30 PM	3:00 PM	Denver 5-6	Admission, Placement, and Outcomes in Higher Education
1:30 PM	3:00 PM	Denver 3	Teacher Education
1:30 PM	3:00 PM	Colorado IJ	Validity, Validation, and Fairness Frameworks and Considerations
1:30 PM	3:00 PM	Penrose 1	Interrogating the Value of Classroom-external Assessment: A Foundational Discussion for the Field
1:30 PM	3:00 PM	Penrose 2	Addressing Items with Multiple-DIF Flags and Other Fairness Issues across Assessment Sectors
3:15 PM	4:15 PM	Colorado IJ	Applications of Large Language Models to Automated Scoring and Feedback
3:15 PM	4:15 PM	Denver 3	Development and Implementation of RTD Task Models for Large Scale Assessment Content Development
3:15 PM	4:15 PM	Denver 4	Within-year Growth Measures and Learning Inferences
3:15 PM	4:15 PM	Colorado CD	Advances in Item Calibration for CAT
3:15 PM	4:15 PM	Penrose 1	Classroom and Formative Assessment
3:15 PM	4:15 PM	Penrose 2	Addressing Equity-Related Educational Challenges
3:15 PM	4:15 PM	Colorado AB	Licensure and Certification
3:15 PM	4:15 PM	Colorado GH	Reliability, Precision, and Generalizability
3:15 PM	4:15 PM	Denver 5-6	Standard Setting
4:30 PM	6:00 PM		Informal Presidential Closing Reception

NEW SESSIONS

195. The Where, What, and How of the Job Market for Measurement Professionals

Organized Discussion
Thursday, April 24
9:45 AM – 11:15 AM
Hilton Denver City Center Lower Level 2: Colorado EF

In this panel discussion including professionals from a wide variety of careers, we'll discuss where measurement experts are employed, what these careers look like, and how one finds and prepares for these careers. This session will provide valuable practical information whether you are a graduate student, early in your career, or considering a new path related to measurement.

Moderator:
Amy Hendrickson, College Board

Presenters:
Jodi Casabianca, BroadMetrics
Wendy Christensen, University of Colorado School of Medicine
Sara Finney, James Madison University
Karoline Jarr, Curriculum Associates
Christoph Scheidiger, Amazon
Zachary Warner, New York State Education Department

196. Considering the Role of NCME in The Current Civil Rights Landscape: A Panel and Discussion Session Organized by the NCME Presidential Task Force on Civil Rights

Saturday, April 26
11:30 AM – 1:00 PM
Hilton Denver City Center Lower Level 2: Colorado EF

In this session, panelists with experience and expertise in education, policy, and civil rights will join NCME leaders and members in conversation about the importance of NCME's voice and responsibility in the current policy and legal context. Implications of changes in the civil rights landscape for measurement in education, and for NCME members themselves, will be discussed, as will opportunities for collaboration and collective action. Audience discussion and interaction with panelists will be interspersed throughout the session.

Moderator:
Britte Haugan Cheng, Menlo Education Research

Presenters:
Michael Feuer, George Washington University
Rochelle Michel, Smarter Balanced
Kimberly Robinson, University of Virginia
Pauline Strong, University of Texas at Austin
Andrew Ho, Harvard University
Amy Hendrickson, College Board
Claudia Ventura, University of Connecticut
Catherina Villafuerte, University of Connecticut

ERRATA

Additions

The following **sessions** have been added to the program.

Date	Time	Room	Title	Speakers
Th, 4/24	9:45 – 11:15 AM	Colorado EF	*The Where, What, and How of the Job Market for Measurement Professionals* (Invited Session)	*Moderator*: A. Hendrickson *Presenters*: J. Casabianca, W. Christensen, S. Finney, K. Jarr, C. Scheidiger, Z. Warner
Sa, 4/26	11:30 AM - 1:00 PM	Colorado EF	*Considering the Role of NCME in The Current Civil Rights Landscape: A Panel and Discussion Session Organized by the NCME Presidential Task Force on Civil Rights* (Invited Session)	*Moderator*: B. Cheng *Presenters*: M. Feuer, R. Michel, K. Robinson, P. Strong, A. Ho, A, Hendrickson, C. Ventura, & C. Villafuerte

The following **presentations** have been added to the program.

Date	Session	Title	Speakers
Fr, 4/25	122	*Something is Not **Always** Better than Nothing: Measurement in LMICs* [e-Board Session #4]	L. Bahry

Cancellations

The following **sessions** have been withdrawn from the program.

Date	Session	Title	Speakers
Sa, 4/26	153	*Can Engagement and Measurement Precision be Enhanced in NAEP Mathematics through MST?*	Y. Kim et al.

The following **presentations** have been withdrawn from the program.

Date	Session	Title	Speakers
Th, 4/24	083	*Adaptively regulated interim theta bounds for optimizing CAT item selection*	S. Lee
Fr, 4/25	109	*Providing Credit for Non-Credit Certificates, One Institution's Journey*	Kristen N. Hodge-Clark
Fr, 4/25	114	*A Comparison of Latent Variable Approaches to Careless Survey Response Detection*	K. Roberts
Fr, 4/25	136	*A Comprehensive Comparison of Item Selection Methods in CAT for Multidimensional Forced-Choice Measures*	Q. Wang, D. Tu, Y. Cai, J. Wei, & J. Ding
Sa, 4/26	146	*Using Multilevel Modeling to Assess Differential Predictive Validity in Graduate Admissions Tests*	C. Patterson & J. Templin
Sa, 4/26	146	*Socioeconomic achievement gap in writing and assessment moderators: A meta-analysis*	C. A. Castillo

ERRATA

Session Room Assignment Updates

Date	Session	Title	New Room
Th, 4/24	038	*More than Words: Intersectionality in Education Research and Policy*	Denver 4
Th, 4/24	043	*Process Data Analyses to Understand Large-Scale Survey-Based Assessment Results*	Mattie Silks
Th, 4/24	044	*Are Standardized Admissions Tests Fair? If Not, How Might We Make Them Fairer?*	Penrose 2
Th, 4/24	067	*Exploring Effective, Equitable Classroom Assessment to Support Teacher Practice and Student Success*	Denver 4
Th, 4/24	068	*Testing Standards Revision: Focus on Foundations, Operations, & Psychometrics (Joint NCME/AERA Session)* ***Note: this session will still take place on Th, 4/24, but is now scheduled for 1:45 – 3:15 PM at the Colorado Convention Center.***	Mile High Ballroom 1EF (Ballroom Level)
Sa, 4/26	170	*DCM for Longitudinal Data*	Mattie Silks
Sa, 4/26	172	*The Future of K12 Assessment & Accountability: Can Recent Research inform policy changes?*	Denver 5-6

Author/Presenter Updates

Date	Session	Title	Speaker (Change)
Th, 4/24	047	*The Developmental Curve of Students' Social-Emotional Competencies and COVID's Impact*	S. Hu (new author, presenter)
Th, 4/24	082	*Direct Estimation: A Flexible Approach for Large-Scale Assessment Data with Plausible Values*	P. Bailey, T. Zhang, E. Sikali (new authors)
Fr, 4/25	091	*Findings from Culturally Responsive Assessments Comprising Original and Adapted NAEP Items*	R. Lopez, J. Sparks, and S. Pillarisetti (new authors)
Fr, 4/25	122	*An English Proficiency Test with Spanish Directions: Helpful or Irrelevant?*	E. Crespo Cruz (new presenter, new author)
Fr, 4/25	132	*Students' COR task solving in economics: time-to-draft with and without ChatGPT*	P. Hodes (new presenter)
Sa, 4/26	139	*Scaling, Linking, and Equating with Test Redesign* [Organized Discussion]	W. Cui (withdrawn panelist)

Chair/Discussant Updates

Date	Session	Title	Chair/Discussant (Change)
Th, 4/24	047	*Social-Emotional Learning Measurements in Service of ALL Students: Lessons Learned around COVID*	S. Hu (new chair)
Th, 4/24	053	*Measurement Invariance and Differential Item Functioning*	A. Nasyrova (new chair)
Fr, 4/25	089	*NCME Career Award*	C. Ibuot (new discussant)
Fr, 4/25	091	*Evaluating the Psychometric Impacts of Cultural Representation in Item Contexts*	M. Rodriguez (new discussant)

FULL SCHEDULE
WEDNESDAY, APRIL 23

For the most up-to-date program information, please visit:
https://www.xcdsystem.com/ncme/program/La91epP/

001. A Brief Introduction to Big Data Techniques Using International Large-Scale Assessment Data

Training Session
8:00 AM - 12:00 PM
Hilton Denver City Center Lower Level 1: Penrose 1

The best definition of big data is that the management and analysis of the data benefits from nontraditional methods. Users may think they need a larger computer, and that will work for a bit. But eventually you need to rethink how you clean and analyze your data to get the most out of scarce computing resources. While educational measurement and psychometrics programs have not traditionally provided any training on big data methods, such as data sharding or distributing computing,that is the topic of this professional development. Using as example the publicly available data for all PISA cycles, attendees will be introduced to basic concepts of big data management and analysis though a mix of theoretical presentation and hands-on analyses. The seminar will focus on how to efficiently manage datasets with millions of rows and hundreds of columns, out of memory analysis, and cluster computing using Spark. At the end of the seminar attendees will be able to productively conduct analysis using Spark parallelization in either personal laptops or high efficiency clusters. The basic concepts and skills acquired in this seminar will provide an effective entry point into Big Data methods.

Presenters:
Tiago Caliço, Organization for Economic Cooperation and Development
Paul Bailey, American Institutes For Research

002. An Overview of Operational Psychometric Work in Real World

Training Session
8:00 AM - 12:00 PM
Hilton Denver City Center Lower Level 2: Colorado AB

This training session will present an overview of the psychometric work routinely done at various testing organizations. The training session will focus on the following topics: (1) outline of operational psychometric activities across different testing companies, (2) overview and hands-on activities to review item and test analyses output, (3) overview and hands-on activities to review equating output, and (4) an overview of data forensic analyses (a.k.a., statistical analyses for test security) to identify test takers who may have engaged in fraudulent behavior. We are hoping that through this training session, participants will get a glimpse of the entire operational cycle, as well as gain some understanding of the challenges and practical constraints that psychometricians face at testing organizations. It is targeted toward advanced graduate students who are majoring in psychometrics and seeking a job in a testing organization and new measurement professionals who are interested in an overview of the entire operational testing cycle. Representatives from various testing organizations with extensive psychometric experience will present on topics related to different processes within an operational cycle.

Presenters:
Hyeonjoo Oh, Riverside Insights
JongPil Kim, Riverside Insights
Jinghua Liu, The National Board of Osteopathic Medical Examiners
Sarah Quesen, WestEd
Hanwook Yoo, Ascend Learning

003. Applications of Bayesian Networks in the Age of AI

Training Session
8:00 AM - 12:00 PM
Hilton Denver City Center Lower Level 2: Denver 4

In the current digital and AI age, information from all directions is overwhelming. It is especially important that we making inferences from evidence based on well-evaluated data from appropriate sources. 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. This training course starts with assessment framework, design, Bayesian network, model building and validation, and applications of Bayesian networks including learning progression (using Netica).

Presenters:
Duanli Yan, ETS
Russell Almond, Florida State University
Diego Zapata-Rivera, ETS
Robert Mislevy, University of Maryland

004. Applying Data Mining Methods to Detect Test Fraud

Training Session
8:00 AM - 12:00 PM
Hilton Denver City Center Lower Level 1: Penrose 2

This session will provide the audience with systematic training on applying various data mining models using software programs like R and/or Python to detect fraud in different test formats, such as computer-based, computer-adaptive, or multistage settings. 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 demonstrations. Further, the advantages and disadvantages of using each software program will be discussed.

This session consists of lectures, demonstrations, and hands-on activities 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 laptops and download the software programs free online. It is expected that attendees will master the basics of specifying various data mining models and applying these models to detect aberrantly behaved test-takers; further, they can apply the skills to their research and datasets.

Presenters:
Kaiwen Man, University of Alabama
Sarah Toton, Caveon
Kylie Gorney, Michigan State University
Qipeng Chen, University of Alabama

005. Best Practices for Artificial-Intelligence Scoring

Training Session
8:00 AM - 12:00 PM
Hilton Denver City Center Lower Level 2: Colorado CD

Increasingly, artificial intelligence (AI) algorithms trained to predict human ratings are used to score constructed responses. Decisions must be made as to whether the AI predictions are adequate to serve as scores. With rapid expansion of AI scoring, including the use of generative AI, there is a need to carefully rethink how we determine if AI produces valid, fair, and reliable scores. The training session will address how to make this judgement. We will discuss evaluating performance of AI algorithms and resulting scores in the context of the larger validity argument for constructed responses and the entire assessment. We will discuss the role of validity evidence for human ratings and the importance of the quality of those ratings for the validity of AI scores. We will also discuss how validity evidence for scores from large language models differs from traditional feature-based models. We will identify essential statistical analyses for evaluating AI scores and how to judge their results. The course is targeted for psychometricians responsible for designing assessments and ensuring scores support the test's claims. It should be of interest to decision-makers and more broadly for anyone working to ensure ethical and responsible use of AI in assessment.

Presenters:
Daniel McCaffrey, ETS
Jodi Casabianca, ETS

006. Build Your Personal Website With Quarto

Training Session
8:00 AM - 10:00 AM
Hilton Denver City Center Lower Level 2: Denver 3

Are you interested in sharing your work, such as scholarship, creative work, and other projects, and increasing your online visibility? Creating a personal website allows you to highlight yourself and your work, develop an online presence that connects people to you and your work, and share your thoughts in a blog or other pages. Having a personal website is a great way for consultants or speakers to market their services and develop a portfolio of past work; for researchers and scholars to share their publications, including archiving for green open access; for public intellectuals to write blog posts and direct readers to their other writings; and for job-seekers to highlight a portfolio of their work for potential employers. This mini training session will provide a basic foundation to get a personal website started and develop the skills to continue building and maintaining your site. We'll leverage Quarto and GitHub Pages to build a website: Quarto makes publishing websites simple while providing a lot of customizability for users interested in going deeper, and GitHub Pages allows users to publish their websites for free.

Presenter:
Matthew Johnson, University of South Florida

007. Comprehensive Statistical Model Evaluation

Training Session
8:00 AM - 12:00 PM
Hilton Denver City Center Lower Level 2: Colorado IJ

Statistical models play a central role in scientific analysis, inference, and decision-making, so it is imperative that researchers diligently and thoroughly evaluate their models before disseminating them. This training session offers an immersive exploration of three perspectives on statistical model evaluation. Attendees will gain a theoretical and methodological understanding of (1) traditional goodness-of-fit testing and bootstrapping procedures, (2) Bayesian prior and posterior predictive model checking, and (3) information-theoretic techniques that adhere to the principle of minimum description length. This discussion will culminate in a simple framework that integrates all three perspectives. Session leaders will then demonstrate a user-friendly Shiny software application that allows users to upload data, specify a statistical model, select any or all of the methods within the framework, and generate a customized model evaluation report. Attendees should bring their own laptops, with R and RStudio installed prior to the session.

The intended audience comprises researchers who use statistical models of any form. The methods covered in this session have been applied to item response theory, factor analysis, and structural equation models, and could be extended to other modeling frameworks. There are no prerequisites for this session, other than general familiarity with the practice of statistical modeling.

Presenters:
Wes Bonifay, University of Missouri
Sonja Winter, University of Missouri
Brian Keller, University of Missouri

008. Data Visualizations and Dashboards to Inform Assessment Development Using R

Training Session
8:00 AM - 12:00 PM
Hilton Denver City Center Lower Level 1: Mattie Silks

Data dashboards in assessment are frequently targeted at end users to understand student performance (Verbert et al., 2013). However, data dashboards can also be a valuable tool in assessment development. Based on lessons learned from the Measures for Early Success Initiative, this training session will teach participants to create data dashboard for psychometric validation. Participants will the learn to create clear and impactful data visualizations using R, specifically focusing on the functionality of the ggplot2 package and Quarto documents. The first half of the session will introduce participants to the principles and techniques for generating intuitive visualizations from early-stage pilot assessment data using the ggplot2 R package. The second half will extend these skills by demonstrating how to integrate multiple visualizations into a cohesive dashboard using Quarto documents, the latest iteration of R Markdown designed for enhanced flexibility and professional presentation. This session is targeted at individuals with a foundational knowledge of R. Attendees are required to bring their own laptops with RStudio pre-installed. By the end of this session, participants will have the capability to effectively communicate complex data to stakeholders with varying levels of data literacy.

Presenters:
Victor Porcelli, MDRC
Electra Small, MDRC
Emily Swinth, MDRC
Rebecca Salzer, MDRC

009. Formative Assessments or the Formative Assessment Process – The -s Makes the Difference
Training Session
8:00 AM - 10:00 AM
Hilton Denver City Center Lower Level 2: Denver 1-2

"Formative assessments" and "formative assessment" differ by one letter, but the differences are profound. The former (with the ""-s"") are periodic summative assessments of learning designed to gauge students' mastery of academic content, before comprehensive summative assessments are taken. More accurately, they are ""interim assessments."" The latter (without the ""-s"") is a process of instructionally embedded practices meant to fundamentally change how students learn academic content, guided by teachers. Research shows it can increase learning and achievement of all students, especially low-performing ones. CCSSO provided one widely used definition: ""Formative assessment is a planned, ongoing process used by all students and teachers during learning and teaching to elicit and use evidence of student learning to improve student understanding of intended disciplinary learning outcomes and supports students to become more self-directed learners."" (CCSSO, 2018) Formative assessment is:

Pre-planned by teachers
Used by teachers and student during instruction and learning
Based on key learning targets
Continuous gathering of evidence of student understanding
Descriptive feedback from the teacher, peers, or themselves is provided
Students adjust their learning tactics as needed
This session will provide an overview of formative assessment.

Presenters:
Edward Roeber, Michigan Assessment Consortium
Tara Kintz, Michigan Assessment Consortium
John Lane, Michigan Assessment Consortium
Ellen Vorenkamp, Michigan Assessment Consortium
Mary Helen Diegel, Michigan Department of Education

010. VEMIRT: An R Package for Conducting High-Dimensional IRT Analysis
Training Session
8:00 AM - 12:00 PM
Hilton Denver City Center Lower Level 2: Denver 5-6

In psychological and behavioral science, the increasing availability of rich survey data and the emerging need to assess multifaceted constructs pose great challenges to existing techniques used to handle and analyze heterogeneous assessment data. The multiple-group multidimensional IRT (MIRT) model that relaxes strict measurement invariance assumption is a viable psychometric tool to establish commensurate measures for the constructs of interest. To calibrate this model more efficiently, we developed a family of innovative Gaussian variational expectation-maximization (GVEM) methods and implemented the methods in the VEMIRT software. This software package comes in two forms: an R package and a Shiny App. In this half-day training workshop, we will introduce MIRT and its role in the analysis workflow for assessment and demonstrate how VEMIRT can assist with all steps of the MIRT applications. The application includes evaluating the dimensionality of the assessment, conducting exploratory and confirmatory item factor analysis, and evaluating measurement invariance across multiple covariates. Workshop attendees will learn about the full spectrum of MIRT theory and applications.

Presenters:
Chun Wang, University of Washington
Weicong Lyu, University of Macau
Gongjun Xu, University of Michigan

012. Harder than Rocket Science? Explaining Assessment to Everyday Audiences

Training Session
10:30 AM - 12:30 PM
Hilton Denver City Center Lower Level 2: Denver 1-2

Measurement professionals must be excellent communicators. Whether launching presentations at school board meetings, propelling stakeholder convenings, or boosting assessment literacy, our messages are the fuel that powers the mission. This workshop gives participants written and visual communication tools to ignite action and benefit society. All measurement professionals looking to add communication strategies to their assessment launchpad are welcome. As professionals, we must be clear in our message to improve, and effect change in educational assessment. Decision makers rely on the data we share and depend not only on the accuracy of our results but the precision of our communication. Nuances often make the difference between test users who can act upon information and those who are left in the dark.

Using case studies, peer collaboration, and direct learning, participants will leave the session with:
Three reasons for the importance of communicating clearly with lay audiences
Framework for evaluating the effectiveness of communication for non-technical audiences
Useful techniques for de-jargoning communication while maintaining credibility
Different personas to consider when developing a communication plan
Techniques to use when creating graphics and pictures for communication

Presenters:
Karoline Jarr
Ashleigh Crabtree, Curriculum Associates

013. Building Monte Carlo Simulations in R

Training Session
1:00 PM - 5:00 PM
Hilton Denver City Center Lower Level 2: Denver 5-6

In this course we will learn how to write Monte Carlo simulations in R. Monte Carlo simulations are an essential tool for both for small-scale investigations and for formal methodological research. Simulation can assess, for example, how much re-testing biases a regression discontinuity design due to undermining the measurement properties of a test score. Our focus is on the best practices of simulation design. Overall, we will show how a specific simulation framework allows for rapid exploration of the impact of different design choices, measurement qualities, and data concerns, and show how simulation can answer questions that are hard to answer using direct computation. For example, available algebraic formulas are often based on asymptotic approximations, which might not "kick in" if sample sizes are moderate; simulation can provide an answer.

In this session we will work through a case study that showcases a modular programming approach for good design. Code, along with an on-line open-source textbook, will be provided and demonstrated. Students should bring laptops, with R and RStudio installed, to follow along. By the end, students will be able to adapt provided code to their own purposes moving forward. Some familiarity with R is assumed.

Presenter:
Luke Miratrix, Harvard Graduate School of Education

014. Cloud Computing and Artificial Intelligence in Psychometric Applications

Training Session
1:00 PM - 5:00 PM
Hilton Denver City Center Lower Level 2: Colorado IJ

Cloud computing has become increasingly popular over the past two decades and has become the cornerstone for AI applications. As practitioners who handle assessment data and do various computing tasks daily, it can be helpful to explore how cloud computing technology can be leveraged to improve efficiency and provide effective solutions to existing daunting challenges. In this workshop, we will cover several AWS core services by (1) conducting machine learning models (ML)/large language models (LLMs)/Generative AI (GenAI) analysis and (2) storing results in a secure S3 bucket and the database on the cloud. Participants do NOT need any prior AWS experience, nor do they need to open an AWS account; we will provide the training materials and hands-on activities through pre-built AWS labs. Upon completion, they will be able to streamline ML/LLM analysis on cloud with storing the results in a secure S3 bucket and the database using AWS services. This is a hands-on training and participants are strongly encouraged to bring their laptops to follow along in order to optimize the learning outcomes from this four-hour training.

Presenters:
Huijuan Meng, Amazon Web Services
Ye Ma
Vinita Talreja
Christopher Taylor, Amazon Web Services
Mingqin Zhang
Ze Wang

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

Training Session
1:00 PM - 5:00 PM
Hilton Denver City Center Lower Level 2: Colorado AB

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, The University of Hong Kong
Wenchao Ma, University of Minnesota
Sangbeak Ye, Florida Atlantic University

016. Incorporating Process Data in Large-Scale Assessments: Methods and Case Studies

Training Session
1:00 PM - 5:00 PM
Hilton Denver City Center Lower Level 2: Denver 3

This training session introduces the fundamental structure and analytic methods of process data, aiming to support the incorporation of response process information in large-scale assessments. The training will combine theoretical instructions on methodologies to analyze process data through sequence-based and feature-based approaches, along with programming code demonstrations, and case study illustrations using process data in PIAAC and NAEP. Specifically, the presenters will focus on four subtopics: (1) creating and selecting informative features from process data in large-scale assessments, (2) extracting and selecting gram-based features from clickstream sequences, (3) computing sequence distance to identify pairwise sequence similarity and conduct sequence clustering, and (4) identifying and interpreting meaningful patterns from process data.

During the four-hour training session, participants will receive an overview of process data structure in large-scale assessments, learn about various approaches to analyzing process data with sequence mining and observed and/or latent feature-based methods, and will broaden their research perspectives on shaping research questions with the assistance of process data in educational assessments through shared examples. The intended audience includes researchers, students, and practitioners with basic knowledge of process data, who are interested in learning or applying data-driven methods to process data analysis. Laptop is recommended but not compulsory.

Presenters:
Qiwei He, Georgetown University
Susu Zhang, University of Illinois at Urbana-Champaign
Elizabeth Tighe, Georgia State University
Xin Wei, Digital Promise

017. Introduction to AI-based Automated Item Generation and Automated Scoring

Training Session
1:00 PM - 5:00 PM
Hilton Denver City Center Lower Level 1: Penrose 2

In the era of artificial intelligence (AI), the field of educational testing faces significant challenges, particularly in test development and scoring, two key innovations include automated item generation (AIG) and automated scoring (AS). Only recently that generative AI has facilitated the development of complex test items on a large scale.

We introduce "the item factory", for managing large-scale test development including automation of item generation, quality review, quality assurance, and crowdsourcing techniques in adaptive testing. We present an overview of the latest natural language processing (NLP) techniques and large language models for AIG, alongside psychometric principles and practices for test development. We discuss the application of engineering principles in designing efficient item production processes (Luecht, 2008; Dede et al, 2018; von Davier, 2017). As AS becomes becomes an integral part of the assessment landscape due to their advantages in reporting time, cost, objectivity, consistency, transparency, and feedback. We aim to demystify AS and provide a comprehensive understanding of its workings. We offer an overview of the design, development, evaluation, and quality control of automated scoring systems, along with practical advice and considerations for practitioners on the applications of these systems into formative and summative assessments (Yan, Rupp, & Foltz, 2020).

Presenters:
Duanli Yan, ETS
Alina von Davier, Duolingo

018. Modeling Item-Level Heterogeneous Treatment Effects: A Tutorial in R

Training Session
1:00 PM - 3:00 PM
Hilton Denver City Center Lower Level 1: Penrose 1

Program evaluators typically examine total test scores as outcomes of educational interventions. However, a recent body of scholarship has demonstrated that interventions can have unique impacts on each item of an outcome measure. These models for "Item-Level Heterogeneous Treatment Effects" (IL-HTE) provide both statistical and substantive insights into the impacts of educational interventions and have implications for critical issues in educational research such as learning transfer, alignment, score inflation, and fade out. In this training session, we will review the logic behind IL-HTE models and learn how to estimate them in R using real data from randomized controlled trials in education research. Knowledge of item response theory, multilevel models, and causal inference in randomized trials is assumed. Attendees should bring laptops with R and RStudio installed. Prior experience with R is helpful but not necessary.

Presenter:
Joshua Gilbert

019. Opening the Black Box: Understanding Content Development for Large Scale Standardized Assessment

Training Session
1:00 PM - 5:00 PM
Hilton Denver City Center Lower Level 2: Denver 4

Instrumentalization is perhaps the most important foundational competency in educational measurement, as it determines the quality of the data that is collected and boundaries of appropriate use(s) of a test. While test design sets many goals and decisions, it is the implementation of that design that ultimately matters most. This training session focuses on that implementation—on content development. Content development occurs between test design and test administration, taking upwards of 30 months and involving many different contributors from different disciplines. Unfortunately, it is not well covered in the literature or in formal educational measurement programs.

While the discipline of psychometrics focuses primarily on quantitative reliability, content development professionals (CDPs) focus primarily on qualitative validity—especially content representation and fairness. This training session lays out a genericized industry standard process used to draft, analyze and iteratively refine items from item writing (i.e., the initial drafts) all the way through psychometric review of field test results and operational test assembly.

Participants will discuss dimensions of item quality, review distractors for how they can subtly alter items, examine sources of relevant item development expertise, and review items using the Item Alignment Examination protocol. Participants should bring their own laptop computers.

Presenters:
Marjorie Wine, Accessible Teaching, Learning and Assessment Systems (ATLAS), University of Kansas
Alexander Hoffman, AleDev Research & Consulting

020. Organizational Leadership for Measurement Experts

Training Session
1:00 PM - 5:00 PM
Hilton Denver City Center Lower Level 2: Denver 1-2

Participants in this training course will learn to identify the skills needed to function effectively as an organizational leader, learn from the real-world experience of measurement professionals who have been effective in leadership roles, and have the opportunity to reflect upon and explore their own interests and skills. Open to all but specifically intended to address the interests of early and mid-career professionals who are engaged in or contemplating organizational leadership roles, individuals from all backgrounds, including those from groups traditionally under-represented in leadership roles, are highly encouraged to participate. The facilitators have both rigorous scientific training and extensive organizational leadership experience. Leveraging NCME's Foundational Competencies in Educational Measurement, dimensions of effective leadership are presented and discussed broadly, informed by real-world experience from within testing organizations and higher education contexts. The training session will be interactive, affording opportunities for self-reflection, small-group activities, and large-group discussions.

Presenters:
Richard Patz, University of California, Berkeley
Jennifer Dunn, Pearson
Michael Rodriguez, University of Minnesota

021. Positionality in Action: Developing Tools to Recognize Bias Across Settings

Training Session
1:00 PM - 3:00 PM
Hilton Denver City Center Lower Level 1: Mattie Silks

The workshop, "Positionality in Action: Developing Tools to Recognize Bias Across Settings,"sponsored and facilitated by the NCME Graduate Student Issues Committee, aims to enhance participants' expertise in crafting impactful positional statements across diverse contexts. Through interactive exercises, attendees will master the art of creating clear, concise, and transparent statements while addressing personal and research biases. Key learning objectives include understanding the essential structure of effective positional statements, applying techniques to articulate viewpoints convincingly, and integrating an awareness of personal bias into research measurement. Participants will engage in practical activities, receive targeted feedback, and apply their newly acquired skills to real-world scenarios. Designed for professionals, academics, and students, this workshop provides actionable strategies for developing well-supported positional statements that critically reflect potential biases. By the end of the session, attendees will be adept at producing transparent and compelling positional statements, significantly enhancing the credibility and impact of their research.

Presenters:
Catherina Villafuerte, University of Connecticut
Nicolas Mireles, James Madison University

Presenters:
Sarah Hammami, Massachusetts Institute of Technology
Claudia Ventura, University of Connecticut
Brein Mosely, Harvard University
Autumn Wild, James Madison University

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

Training Session
1:00 PM - 5:00 PM
Hilton Denver City Center Lower Level 2: Colorado CD

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 all score users can interpret and use the information in appropriate ways. More recently interest around interactive 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 multi-step 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 users in different contexts. In this session, we will focus on parents, teachers/administrators, and policy makers as three focal user 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 components.

Presenters:
Priya Kannan, WestEd
Diego Zapata-Rivera, ETS
Richard Feinberg, NBME
Francis O'Donnell
April Zenisky, University of Massachusetts, Amherst

023. Culturally and Linguistically Responsive Assessment for Learning in STEM

Training Session
3:30 PM - 5:30 PM
Hilton Denver City Center Lower Level 1: Mattie Silks

In this session, we explore culturally and linguistically responsive (CLR) assessment for learning (AfL) to develop inclusive and equitable assessments for diverse populations. The learning objectives are: (a) Understanding how to integrate CLR principles into assessment design to maximize learning for diverse learners; (b) Applying CLR to STEM classrooms and large-scale assessments; (c) Introducing two CLR AfL models—co-design and expanded evidence-centered design to develop assessments that reflect students' cultural and linguistic backgrounds based on real-world, authentic, case studies / scenarios; and (d) Engaging with inclusive assessment strategies for redesigning AfL tools (e.g., rubrics, tasks, and digital content) from a CLR perspective to help ensure fairness and validity. Attendees will engage in interactive discussions, collaborative learning, and share experiences and strategies for implementing CLR AfL in educational settings. Through these learning objectives, attendees will better understand how to implement best practices for inclusive assessment including creating and implementing assessments that are sensitive to cultural and linguistic diversity, with practical strategies for real-world application. This session is intended for educators, assessment professionals, and researchers with a basic understanding of educational assessment. There are no prerequisites for attendance. Attendees should bring their own laptops. No special software is necessary for this session.

Presenters:
Maria Oliveri, Purdue University
Mya Poe, Northeastern University
Archer Harrold, University of Michigan

024. PACER for Statistics and Psychometrics

Training Session
3:30 PM - 5:30 PM
Hilton Denver City Center Lower Level 1: Penrose 1

PACER is a free web-based psychometric platform available at http://www.paceronline.com/. This software was designed to provide a contribution to psychometric society to advance the exploration of data in a simple environment, thus aligning with the theme of this year's NCME conference.

It is an easy-to-use, multi-facted environment that can support IRT item calibrations, generating ability estimates, test equating, regularized regressions (Ridge, Lasso, Elastic Net), automated test assembly (ATA), cognitive diagnostic models, mixed linear models with nested and crossed random effects accounting for measurement error, generating response probabilities, and other basic statistical and psychometric methods.

This session will introduce users to the environment and provide sample data sets for application and practice.

Presenter:
Harold Doran, HumRRO

NCME BOARD OF DIRECTORS MEETING

Members of NCME are invited to attend as observers.
Wednesday, April 23
4:30 PM TO 6:30 PM
Hilton Denver City Center Lower Level 1: Gold Coin

025. Serving Diverse Student Groups Through Accommodation and Accessibility Support in Large-Scale Assessments

Coordinated Paper Session
8:00 AM - 9:30 AM
Hilton Denver City Center Lower Level 2: Colorado IJ

In the evolving landscape of large-scale assessments, ensuring equitable access for diverse student populations remains a significant challenge. This coordinated session delves into the intersection of accommodation, accessibility, and assessment outcomes through five studies that examine the opportunities and challenges of enhancing support for students with disabilities and multilingual learners. The session begins with an examination of accommodation patterns across different subjects and their impact on academic performance among students with disabilities. The second paper investigates the combined use of extended time and separate sessions, two commonly bundled accommodations. The third paper shifts the focus to multilingual learners, exploring their use of bilingual test forms and text-to-speech features. The fourth paper explores students' self-reported use of universal accessibility features and perceptions of their usefulness. The final paper presents an innovative approach to standardized assessment formats, investigating the potential of non-language-based assessments to provide a more equitable measurement of knowledge and skills for English-as-an-additional-language learners. Together, these papers contribute to the ongoing discussion on how to optimize assessment practices to support all learners. The session concludes with a discussion on the ethical implications of using accommodation data, emphasizing the importance of privacy, fairness, and compliance with existing standards.

Session Organizer:
Hyunah Kim, Education Quality and Accountability Office

Chair:
Hyunah Kim, Education Quality and Accountability Office

Presentations:

Accommodation Patterns and Their Association with Academic Performances in Students with Disabilities
Pei-Ying Lin, University of Saskatchewan

Examining Bundled vs. Individual Accommodations of Extended Time and Separate Sessions
Xin Wei, Digital Promise

Students' Choice of Spanish Text and Text-to-Speech on a Low-Stakes Mathematics Test
Heather Buzick, ETS

Universal Accessibility Features: Who Uses Them and Who Finds Them Useful?
Hyunah Kim, Education Quality and Accountability Office; Norah Rayfield, Education Quality and Accountability Office; Jennifer Hove, Education Quality and Accountability Office

Non-Language Based Large-Scale Assessment Formats as Service to Society
Man-Wai Chu, University of Calgary; Marie-Claire Shanahan, University of Calgary; Sammy Yung, University of Calgary; Israt Jahan, University of Calgary; Christopher Hass, Rutgers University; Benjamin Giers; University of Calgary; David Giers, University of Calgary

Discussant:
Cara Laitusis, Center for Assessment

026. Computer-supported Collaborative Social Emotion Learning and Assessment in Classroom

Coordinated Paper Session
8:00 AM - 9:30 AM
Hilton Denver City Center Lower Level 1: Penrose 1

Social and Emotional Learning (SEL) is a vital aspect of education that helps individuals develop healthy identities, manage emotions, build strong relationships, and make responsible decisions. However, teaching and assessing SEL is challenging as it requires practices involving multiple participants, making it difficult to capture evidence through traditional classroom methods. In this coordinated session, we introduce our innovative solution to learning and assessing SEL based on computer-supported collaborative learning system in classroom settings. We will discuss the design, implementation, and how AI and data analytics help to identify evidence from the learning process data to assess SEL by going beyond the self-reported approach that is typical in SEL evaluation. The project is jointly funded by the Education Innovation Research program of the U. S. Department of Education and the ETS Research Institute.

Session Organizer:
Jiangang Hao, ETS

Presentations:
Teaching and Learning SEL: Piloting a Digitally Enhanced Curriculum
Patrick Kyllonen, ETS; Jiangang Hao, ETS; Jessica Andrews-Todd, ETS; Emily Kerzabi, ETS

Designing Digital Enhancements for a SEL Classroom
Emily Kerzabi, ETS, Enruo Guo, ETS; Jiangang Hao, ETS; Jessica Andrews-Todd, ETS; Patrick Kyllonen, ETS

AI and Analytics to Support Computer-supported SEL
Jiangang Hao, ETS; Wenju Cui, ETS; Yang Jiang, ETS; Patrick Kyllonen, ETS; Emily Kerzabi, ETS; Jessica Andrews Todd, ETS

A Collaborative SEL Framework to Provide Evidence for Formative Skill Assessment
Jessica Andrews Todd, ETS; Emily Kerzabi, ETS; Jiangang Hao, ETS; Patrick Kyllonen

Discussant:
Noreen Webb

027. Generative AI in Education: Applied Ethics & Evaluating Model Performance (Joint NCME/AERA Session & SIGIMIE-Organized Session)

Coordinated Paper Session
8:00 AM - 9:30 AM
Hilton Denver City Center Lower Level 2: Denver 1-2

With the explosion of Generative Artificial Intelligence (GenAI) solutions in 2022, there has been intense excitement and activity in applying LLM-powered algorithms in education contexts. Now that we've had a few years to reflect on the potential applications of LLMs and conduct research, what have we learned? This session by NCME's SIG on AI in Measurement and Education (AIME) includes collaborative activities and individual research shared during the year.

In this coordinated session, Okan Bulut from the University of Alberta will present a White Paper about Responsible AI use to address concerns such as validity, reliability, fairness, equity, and test security. Kristen DiCerbo and Lauren Deters from Khan Academy discuss challenges in using large language models (LLMs) for math tutoring and present their work to compare model performance and identify the need for new datasets for validation. Maggie Beiting-Parrish of the Federation of American Scientists examines fairness analysis in AI-based autoscoring research conducted with NAEP assessments, discussing the importance of identifying and mitigating biased results. John Whitmer,

also from the Federation of American Scientists, presents an evidence map for evaluating GenAI's performance in educational contexts, applying traditional measurement principles like validity and reliability to evaluate research studies.

Session Organizer:
John Whitmer, Federation of American Scientists

Chairs:
John Whitmer, Federation of American Scientists
Christopher Ormerod

Presentations:
AI in Educational Measurement: Addressing Challenges and Ensuring Responsible Use
Okan Bulut, University of Alberta

Comparing LLM Evaluation Accuracy in Conversational Interactions
Kristen Dicerbo; Lauren Deters

Testing AI Fairness in NAEP Autoscoring Research & Developing Best Practices
Magdalen Beiting-Parrish; National Center for Education Statistics/ Federation of American Scientists

Mapping Measurement Principles to GenAI Research Studies: An Initial Evidence Map
John Whitmer; Federation of American Scientists

Discussant:
Magdalen Beiting-Parrish, National Center for Education Statistics/ Federation of American Scientists

028. Data-intensive research in educational measurement using the Item Response Warehouse

Coordinated Paper Session
8:00 AM - 9:30 AM
Hilton Denver City Center Lower Level 2: Denver 3

Data from educational measures and other psychometric instruments is ubiquitous but has historically been difficult to reuse given the lack of a common standard and central repository. The Item Response Warehouse (IRW) is an attempt to redress this matter by making such data publicly available in a single location and in a common format. This session features research papers that highlight the range of methodological studies made possible by the IRW. Paper 1 is meant to empirically ground future simulation studies by analyzing the spectrum of item difficulty distributions that are observed empirically. Paper 2 considers the effects of RCTs on the psychological composition of the outcome measure using ideas from the rapidly evolving field of psychometric networks. Papers 3-5 consider the degree to which alternative models—a range of polytomous models, unipolar versus bipolar models, and continuous versus class models—produce qualitatively different views of the educational or psychological phenomena in question. This session is designed to spur new ideas about how educational measurement research can be more informed by data and, to that end, will be accompanied with code resources to allow interested researchers to begin to implement their ideas in IRW data.

Session Organizer:
Benjamin Domingue, Stanford University

Chair:
Benjamin Domingue, Stanford University

Presentations:

Incorporating Characteristics of Real-World Item Difficulty Distributions into Simulation Design
Lijin Zhang, Stanford Graduate School of Education

Estimating Casual Effects on Psychological networks Using Item Response Theory
Joshua Gilbert

Prediction quality of a range of polytomous models applied to a large data corpus
Klint Kanopka, New York University

Evaluating Unipolar Versus Bipolar Representations of Educational and Psychological Constructs
Qi Huang, University of Wisconsin

A comparison of the predictive performance of continuous and dichotomous latent trait models
Wanjing (Anya) Ma; Benjamin Domingue, Stanford University

Discussant:
Chun Wang, University of Washington

029. Item Difficulty Modeling in Support of Test Development and Validation

Coordinated Paper Session
8:00 AM - 9:30 AM
Hilton Denver City Center Lower Level 2: Denver 4

Item difficulty modeling (IDM) seeks to identify relationships between the response demands items and the difficulty of those items. Analyses commonly include predictors of difficulty related to item format, content alignment, cognitive complexity, and linguistic complexity. IDM research potentially supports several goals: improving item development practices, identifying construct-relevant and construct-irrelevant factors associated with difficulty, and evaluating hypotheses about the cognitive demands of items and response processes employed when responding. IDM research dating back to the 1940s has identified valuable predictors of difficulty and evaluated statistical techniques for modeling difficulty. However, IDM remains an academic exercise when research findings are not put to their proposed uses. To encourage the broader application of IDM findings, this session focuses on recent IDM projects whose results are being used to improve the development and validation of large-scale assessments in four testing organizations. In these projects, IDM was applied in a variety of contexts including college admissions, alternate assessments, and adaptive assessments connected to K–12 instructional programs. Presenters will each discuss their modeling framework, important predictors of difficulty, as well as current and future applications to test content development.

Session Organizer:
Jeffrey Steedle

Chair:
Jeffrey Steedle

Presentations:

Evaluating ALDs as a Component of Cognitive Complexity for an Alternate Assessment
Christina Schneider; *Jing Chen; Christy Kulczycki, Cambium Assessment, Inc., Matthew Gill, South Dakota Department of Education*

Item Difficulty Modeling Applications for the ACT
Xin Li, Melinda Taylor

Providing Guidance to Item Writers through Item Difficulty Modeling
Tyler Hefty, Renaissance; Tony Albano, University of California - Davis, Edison Choe, Renaissance; Adam Wyse, Renaissance

Predicting Item Difficulty in K–12 Assessment Using Machine Learning
Mubarak Mojoyinola; Nathan Minchen; Jeffrey Steedle

Discussant:
Shonai, Someshwar, University of North Carolina

030. Ensuring Fairness and Consistency in Assessment Practices

Coordinated Paper Session
8:00 AM - 9:30 AM
Hilton Denver City Center Lower Level 2: Denver 5-6

In alignment with the 2025 NCME theme of "In Service of Society," this session addresses fairness and consistency in various aspects of large-scale and high-stake assessment amidst evolving social contexts through four studies. The first study evaluates two calibration strategies for multidimensional constructs with various hierarchical structures, providing guidelines for computing composite scores and assisting the decision-making process to set cut score. The second study examines the consistency of scale scores across test administrations affected by policy-driven population changes, using empirical data from a large-scale assessment. The third study proposes a holistic differential item functioning (DIF) framework for two large-scale assessments with complex group memberships, offering insights into understanding the hidden sources of DIF. The final study investigates item parameter drift (IPD) for polytomous items, focusing on the impact of differences in each score category and further providing flagging criteria for detecting drifted polytomous items. Collectively, these studies offer practical insights for practitioners navigating the challenges from more complex and innovative testing programs.

Session Organizer:
Yawei Shen

Chair:
Siyu Wan, ABIM

Presentations:
Evaluating Calibration Strategies for Composite Test Scores: Impacts on Classification Accuracy
Yawei Shen; Junhui Liu; Hongyu Diao

Maintaining Score Scale Consistency When the Population Changes
Brad Moulder

A Holistic Approach to Detecting Differential Item Functioning Using Mixture IRT Models
Junhui Liu

Identifying Item Parameter Drift for Polytomous Items in Operational Testing Programs
Yi Yang, ETS; Junhui Liu; Dongwei Wang, American Board of Surgery

Discussant:
Jiawei Xiong, Curriculum Associates

031. Electronic Board Session 1

Electronic Board Session
8:00 AM - 9:30 AM
Hilton Denver City Center Lower Level 2: Colorado GH

Presentations:
1. **Enhancing LSAMP Program Outcomes Through Targeted Non-Cognitive Growth and Research Support**

Jaylene Patterson, University of Kentucky; Kelly Bradley, University of Kentucky; Lynette Michaluk, Western Virginia University
This proposal uses Wright map analysis to identify areas for improvement in the Louis Stokes Alliance for Minority Participation (LSAMP) program. By addressing gaps in research dissemination and support for minority STEM students, this study offers actionable recommendations to enhance LSAMP's effectiveness and bolster student persistence and success.

2. **Explaining Performance Gaps with Problem-Solving Process Data via Latent Class Mediation Analysis**
Sunbeom Kwon, University of Illinois, Urbana-Champaign; Susu Zhang; University of Illinois at Urbana-Champaign
This study aims to explain the performance gaps between groups using problem-solving process data. We propose a latent class mediation analysis identifying the latent classes underlying the distribution of sequence features and assessing the indirect effect of the group membership on the final outcome through the latent class mediator.

3. **The Future of Psychometrics: Integrating Critical Perspectives**
Elizabeth Anderson, Founder, Embraced Wisdom Resource Group; Temple Lovelace, Assessment for Good, AERDF, YJ Kim
This paper examines the hisotry and future of psychometrics, addressing how it has atributed to bias, yet how it can contribute to an asset-based approach to classroom assessment. By integrating theoretical, practical perspectives, and case examples, it explores how a more inclusive approach to assessment can honor diverse learners.

4. **More Than the Sum of its Parts: The Relationship Between Individual and Group Ratings of Culturally Responsive Schooling**
Lisa Keller, University of Massachusetts, Amherst; Stella Kim, University of North Carolina Charlotte
This study challenges the notion that a group rating is equivalent to a composite of individual ratings by analyzing the relationship between group and individual ratings. Using the coefficient of consensus, strategies to explore both summary and dispersion are investigated. Models to incorporate school and stakeholder variables are presented.

5. **Comparison of Lexile and Quantile Measures Across Assessments**
Hanh Nguyen
Lexiles and Quantiles, developed by MetaMetrics, are often reported alongside math and reading assessment scores. This study investigates how similar the Lexile and Quantile results are when comparing across different assessments. Although there are many reasons why differences may occur, helping educators understand and interpret different results remains a challenge.

6. **Identifying At-Risk Readers Using Finite Mixture Models and Decision Trees**
Xuejun Ji; Malt Joshi, Texas A&M University
This study investigates the role of home literacy environments and socioemotional competence in shaping reading achievement. Finite mixture and decision tree models were applied to data from the Early Childhood Longitudinal Study to identify early-grade students at risk of reading difficulties. The findings provide insights into early interventions.

7. **Exploring the Validity of Mathematics Self-Concept: Key Findings from the PISA 2022**
Onur Ramazan, Washington State University; Shenghai Dai, Washington State University; Robert Danielson, Washington State University - Spokane
The present study examined the concurrent, discriminant, and predictive validity of mathematics self-concept (M-SC) within PISA 2022 framework. Results supported M-SC as a significant predictor of mathematics performance, distinct from mathematics self-efficacy and significantly relating to mathematics motivation. Findings suggested M-SC's psychometric robustness as a scale, informing future applied research.

8. **Applying Exploratory Graph Analysis to Assess Differential Item Functioning of Multiple Grouping Variables**
   Chi-Chen Chen, National Academy for Educational Research; Kuan-Yu Jin, Hong Kong Examinations and Assessment Authority
   This study introduces the exploratory graph analysis (EGA) method for assessing differential item functioning (DIF) with multiple grouping variables simultaneously. Through a series of simulation studies, we showed that EGA can provide an intuitive and efficient multiple-group DIF analysis, offering insights into the underlying causes of DIF items.

9. **Interrater Reliability Comparisons with Generalizability Theory and Structural Equation Modeling**
   Jason Immekus; W. Holmes Finch, Ball State University; Brian French, Washington State University
   Interrater reliability estimates can differ between rater pairs. These differences can influence score accuracy. A proposed method grounded in Generalizability Theory and Structural Equation Modeling is compared to the W statistic for reliability coefficient comparisons. The simulation results show better performance for the proposed method compared to W.

10. **Bootstrap Methods for Standard Error Estimation in Multivariate Generalizability Theory**
    Stella Kim, University of North Carolina Charlotte; Sungyeun Kim, Incheon National University
    The purpose of the study is to propose three methods for computing SEs for MGT using bootstrap techniques and discuss the methods in terms of their practical and scholarly implications.

11. **Classification Accuracy: Psychometric Model Approach or Clinical Group Approach?**
    Huan Liu, Riverside Insights; Hyeonjoo Oh, Riverside Insights; Michael Custer, Riverside Insights; JongPil Kim, Riverside Insights
    This study evaluates classification accuracy indices of the Battelle Developmental Inventory, 3rd Edition (BDI-3), using both psychometric model and clinical group approaches. The results indicate that both approaches produced comparable classification accuracy indices across all special groups. Both approaches are recommended in assessing the classification accuracy of clinical tests.

12. **Perceived Fairness and Performance-Based Assessment: College Students' Perceptions of Assessment Methods**
    Sanjay van Buel, Erasmus University Rotterdam; Karen Stegers-Jager, Health Academy Radboudumc; Janneke Oostrom, Tilburg University; Marise Born, Erasmus University Rotterdam
    This study examined the impact of standardization and students' minority status on fairness perceptions of performance-based assessment. Findings showed that structured assessment methods are preferred over more unstructured assessment methods, with ethnic group reporting similar preferences. This has implications for assessments currently used in higher education.

13. **Detecting Learning Effects in Small Test Modules for Longitudinal Assessments**
    Yanlin Jiang, NCCPA; Joshua Goodman, National Commission on Certification of Physician Assistants; Andrew Dallas; Andrzej Kozikowski, NCCPA
    The study proposed an approach to detecting learning effects for the test modules or interested content subdomains with small item sets in a longitudinal assessment. The procedure was demonstrated using high-stakes medical exam data. It provides an option to examine learning effects when commonly used models/indices are not feasible.

14. **Extending the Use of Consequential Validity Ratio for Binary Criterion Measures**
    Yixiao Dong, University of Denver; Kushmakar Baral, University of Denver; Denis Dumas, University of Georgia
    The Consequential Validity Ratio (CVR) is a method for quantifying how well test scores avoid the improper influence of participant demographics in predicting a validation criterion. The original CVR can only be used with the continuous criteria, and the present study extends its use for binary criterion measures.

044. Are Standardized Admissions Tests Fair? If Not, How Might We Make Them Fairer?

Organized Discussion
8:00 AM - 9:30 AM
Hilton Denver City Center Lower Level 1: Penrose 2

The fairness of standardized tests has long been questioned by those outside the measurement field (e.g., Gould, 1981; Nairn, 1980). Due to an increased awareness of the imperative for social justice, dramatic changes in national demographics, and other factors, that questioning has more recently been taken up by members of the educational measurement community itself (e.g., Bennett, 2023; Faulkner-Bond, 2022; Lyons, Johnson, & Hinds, 2021; Russell, 2023, 2024; Sireci, 2020; Walker et al., 2023). Although other factors also contribute, fairness issues are cited as a cause of the postpandemic rise of test optional policies and the adoption of blind policies by all public institutions in California (FairTest, 2024; Nietzel, 2021). Even so, some highly selective institutions have recently reinstated their admissions testing requirements (Diep, 2024; McLean, 2024), in part due to concerns that alternative measures like grades, course rigor, recommendations, and personal essays may not be as effective or as fair as tests (Rosinger, Ford, & Choi, 2021). This organized discussion will explore the meaning of "fairness" in the admissions testing context, ideas about how to make tests fairer, and how we might productively move forward given divergent points of view.

Chair:
Randy Bennett, ETS

Presenters:
Randy Bennett, ETS
Li Cai, University of California - Los Angeles
Kurt Geisinger
Rebecca Zwick, University of California, Santa Barbara
Malik Boykin, Brown University

032. DIF Applications

Individual Paper Session
8:00 AM - 9:30 AM
Hilton Denver City Center Lower Level 2: Colorado AB

Chair:
Julian Siebert, Stanford University

Presentations:

Model-Based Detection of Passage Bias in Oral Reading Fluency Assessments
Gozde Sirganci, Southern Methodist University; Yusuf Kara, University of Miami; Joseph F. T. Nese, University of Oregon, Akihito Kamata
This study proposes a model-based approach for detecting DPF in ORF assessments, initially using measurement invariance testing. We use CFA parameterization of a measurement model defined by continuous WCPM scores to indicate latent ORF ability. The findings aim to enhance ORF assessments, reducing construct-irrelevant variance and promoting fairer passage selection.

Differential Grading Standards: Evidence from HSTS & NAEP 2019 Mathematics Data
Darrick Yee, American Institutes for Research; Winifred Wilberforce, Ohio State University
Differential grading standards have been discussed for many decades but rarely investigated at the national level. We link student transcript data from the High School Transcript Study of 2019 to NAEP 2019 mathematics item responses to estimate differential grading among student subgroups using an item response theory approach.

DIF Analysis in Large Scale Assessments. A Case Study in SABER 11
Víctor H Cervantes, University of Illinois Urbana-Champaign; John Alexander, Calderón Instituto Colombiano para la Evaluación de la Educación - ICFES; Nelson Andrés Rodríguez, Instituto Colombiano para la Evaluación de la Educación - ICFES
We present a set of simulation studies that help guide decision-making with respect to DIF analysis in Colombia's SABER 11 and that may be useful for the analyses of DIF in similar LSAs. Results support the validity and fairness of score comparisons of the Mathematics test forms of SABER 11.

Incorporating Measurement Properties When Studying Demographic Disparities in Exam Performance
Siyuan Chen, The Ohio State University; Andrew Heckler, The Ohio State University, Jolynn Pek, The Ohio State University
There is evidence that students from different demographic backgrounds receive different exam scores. Less studied is whether this difference in scores reflects discrepancies in student abilities relative to measurement bias. This study applies a novel method to evaluate the distinct effects of demographic variables on exam items and performance levels.

Discussant:
Christiana Akande, National Commission on Certification of Physician Assistants

033. Student Growth Modeling

Individual Paper Session
8:00 AM - 9:30 AM
Hilton Denver City Center Lower Level 1: Mattie Silks

Chair:
Yeow Meng Thum

Presentations:

Skill Mastery Profiles and Academic Growth
Jinah Choi, Edmentum; Sonya Powers
This study analyzes the relationship between students' pass rate profiles in skill mastery quizzes,

academic growth in overall assessments, and their learning volume over time. The goal of this research is to identify the optimal level of difficulty for a supplemental curriculum that best promotes learning.

Measurement of growth in the context of Alternate English language proficiency assessment
Yun-Kyung Kim, UCLA/CRESST; Eunhee Keum, CRESST; Li Cai, University of California - Los Angeles
This study proposes two longitudinal item response theory models that address systematic missingness often found in Alternate English language proficiency assessments. A simulation study will demonstrate how the proposed models improve the accuracy of student growth percentile estimates, and an empirical example will highlight the usage of the growth estimates.

Growth Modeling for Bayesian Priors in Repeated Measurement of Student Ability
Logan Rome, Curriculum Associates; Sanford Student, University of Delaware; Ted Daisher
When students are measured repeatedly over time, Bayesian ability estimation methods may leverage collateral information, such as previous scores, to predict the prior distribution. We explored a longitudinal growth modeling approach to estimating the prior distribution, and the effect of iteratively fitting the growth model on student ability estimation.

Analyzing Student Performance Using Multilevel Modeling
Anh Thu Le, University of Notre Dame; Doris Luft Baker, The University of Texas at Austin
This study uses multilevel modeling to assess how beginning-of-year performance, conditions, and their interaction affect end-of-year performance for Spanish-English bilingual second graders across 36 schools and two cohorts. Finding suggested a positive impact of beginning-of-year performance and a significant negative interaction with conditions, highlighting the importance of early assessments.

Discussant:
Yeow Meng Thum

034. New IRT Statistical Tools

Individual Paper Session
8:00 AM - 9:30 AM
Hilton Denver City Center Lower Level 2: Colorado CD

Chair:
Jinwen Luo, UCLA

Presentations:
Identifying Ordinal Factor Analysis Models via Integer Coding
Ed Merkle, University of Missouri; Sonja Winter, University of Missouri; Ellen Fitzsimmons, University of Missouri
We propose new identification constraints for ordinal factor analysis. The constraints place the latent variable on a scale from 1 to K, where K is the number of ordered categories. The constraints make models more intuitive, because researchers are already accustomed to treating ordinal variables as continuous.

An Evaluation of IRT Characteristic Curve Linking Methods with Outlying Common Items
Shaojie Wang, Guangdong University of Education
This study examined the impact of outlying common items on the performance of IRT characteristic curve linking methods. Results indicated that the information-weighted characteristic curve methods outperformed traditional methods, maintaining these advantages across various simulated conditions. Additionally, all IRT linking methods exhibited high fidelity in recovering classification accuracy and consistency.

Information Functions of Model-based Oral Reading Fluency
Xin Qiao, University of South Florida; Cornelis Potgieter, Texas Christian University; Akihito Kamata Paul, Foster Southern Methodist University
The computer-based oral reading fluency (ORF) assessments are important tools for identifying at-risk students in various education settings. We provide derivations and demonstrations of information functions based on an ORF model for estimating latent variables and deriving words read correctly per minute (WCPM) scores.

Discussant:
Sijia Huang, Indiana University Bloomington

035. Connecting Performance Level Descriptors and Released Items to Improve Understanding and Communication (SIGIMIE-Organized Session)

Organized Discussion
8:00 AM - 9:30 AM
Hilton Denver City Center Lower Level 2: Colorado EF

We can all agree that it is imperative to understand what students know and can do. Yet, communicating what proficiency means in terms of learning standards or other measured constructs remains a challenge. This session describes how released test items linked to performance categories can improve the understanding of what students know and can do for end-users of assessment results. Presenters will provide a foundational overview for three applications. Attendees will have the opportunity to share successes, and challenges in this work, or propose questions related to the topics during the discussion. Information provided will include how the well-articulated performance level descriptions (PLDs) and released items mapped to those descriptions can be used to (1) train teachers to support and improve instruction, (2) support score users in understanding what "proficiency" means in terms of learning standards, and (3) extend the performance level descriptions and connections to other policy movements. This session will be of interest to those involved with the users of assessment results as well as to those who design, implement, and evaluate district, state, interim, or other large-scale assessments including state education agencies, policymakers, and other partner groups from research to classroom practice.

Chair:
Regina Lewis, Maine Department of Education

Moderator:
Regina Lewis, Maine Department of Education

Presenters:
Bryan Drost, Rocky River City School District
Melissa Fincher, edCount, LLC; Ellen Forte, edCount, LLC
Zachary Warner, New York State Education Department

036. A Framework to Guide Development and Evaluation of AI-Based Formative Assessment Systems

Coordinated Paper Session
9:45 AM - 11:15 AM
Hilton Denver City Center Lower Level 2: Denver 1-2

In the age of the AI app gold rush, the proliferation of AI-driven formative assessment systems necessitates the establishment of clear standards for their development and evaluation. This coordinated session aims to address the urgent need for comprehensive guidelines that align with Black and Wiliam's theory of formative assessment, theories of learning and instruction, and established standards such as those from AERA, APA, NCME, and ITC/ATP. The session will begin

with an introduction to the broader issues and the importance of these standards. It will include four papers: the first presenting a comprehensive framework for developing and evaluating AI-based formative assessment systems, and the subsequent three applying this framework to different AI-based formative assessments. The session will conclude with discussant remarks and a Q&A segment, fostering a deeper understanding and dialogue among attendees.

Session Organizer:
Joshua Wilson, University of Delaware

Presentations:
A Framework to Guide Development and Evaluation of AI-Based Formative Assessment Systems
Joshua Wilson, University of Delaware; Paul Deane, ETS; Christopher Ormerod, Corey Palermo, Measurement Incorporated; Mark Shermis, Performance Assessment Analytics, LLC

Designing an AI-Based Formative Assessment
Paul Deane, ETS; Evaluating Cambi

Evaluating Cambi
Christopher Ormerod; Amy Burkhardt, Cambium Assessment; Susan Lottridge, Cambium Assessment

Fostering Self-Regulated Learning in Writing: A Standards-Based Analysis of MI Write
Corey Palermo, Measurement Incorporated

Discussant:
Mark Shermis, Performance Assessment Analytics, LLC

037. Enhancements to the ACT: A Summary of Research Evaluating Use of ACT Scores

Coordinated Paper Session
9:45 AM - 11:15 AM
Hilton Denver City Center Lower Level 2: Colorado AB

The enhanced ACT will be launched in spring 2025 and ongoing research is planned to document evidence evaluating interpretations and uses of ACT scores, with a focus on whether the strength of evidence is different for the new/enhanced ACT relative to the old/traditional ACT. In this session, we will describe what is changing for the ACT and present rationale for the changes. We will then present results from ongoing research focusing on various aspects of validity (content, cognitive process, internal structure, concurrent, and predictive) and measurement properties such as reliability, score comparability, and timing. We will also present student perceptions of the enhanced ACT relative to the traditional ACT.

Session Organizer:
Jeff Allen, ACT, Inc.

Chair:
Joanna Gorin, ACT Education Corp.

Presentations:
Introduction to ACT Enhancements
Joanna Gorin, ACT Education Corp.

Evidence of Content Validity and Cognitive Process Validity
Jay Thomas, ACT Education Corp; Stacy Dreyer, ACT Education Corp; Scott Johanningmeier, ACT Education Corp

Impact of ACT Enhancements on Student Perceptions and Performance
Shalini Kapoor; Chi-Yu Huang, ACT; Hongling Wang, ACT; Chen Qiu, ACT; Scott Wood, ACT, Inc.; Dongmei Li

Construct Equivalency, Score Comparability, and Psychometric Properties of the Enhanced ACT Test
Dongmei Li; Ann Arthur, ACT Education Corp.; YoungWoo Cho, ACT; Chi-Yu Huang, ACT; Shalini Kapoor

Enhanced ACT Predictive Validity and Composite Score Comparability
Jeff Allen, ACT, Inc.; Tyler Cruce, ACT

Discussant:
Brian Gong, Center for Assessment

038. More than Words: Intersectionality in Education Research and Policy

Coordinated Paper Session
9:45 AM - 11:15 AM
Hilton Denver City Center Lower Level 2: Colorado EF

This session explores how to better incorporate intersectionality into education research and policy to promote fairness and inclusivity, aiming to provide practical information for applied researchers. The first paper discusses how traditional psychometric methods often fail to capture the diverse experiences of test-takers with multiple intersecting identities, advocating for approaches that enhance validity and fairness. The second paper critiques the superficial use of intersectionality in policy, revealing the effects of performative allyship and proposing strategies for genuinely culturally responsive practices. The third paper calls for deeper theoretical engagement with intersectionality in quantitative research to avoid misuse of the concept. The fourth paper challenges deficit-based views of bilingual learners, promoting assessment practices that recognize and leverage their linguistic strengths. The fifth paper examines the limitations of traditional bias detection methods given evolving demographic data standards, suggesting new approaches to address the complexity of intersectional data. Together, these papers urge researchers to move beyond surface-level commitments and outdated methods, advocating for robust, equity-driven practices that fully embrace the complexity and diversity of our students.

Session Organizer:
Sarah Quesen, WestEd

Chair:
Sarah Quesen, WestEd

Presentations:
Psychometric Methods for Intersectional Fairness: Navigating Standardization and Individualization
Tony Albano, University of California - Davis

Educational Policy Perspectives on Intersectionality and Performative Allyship
Lauren White; Jaylin Nesbitt

A Call for Theoretical Emphasis in Intersectional Analysis
Olivia Szendey; Michael Russel, Boston College

From Deficit to Strength: Shifting Perspectives on Spanish-English Bilinguals in KEAs through Bilingual Profiles
Andrea Rolla, Gabriela Mottesi, WestEd; Mel Wylen, WestEd

Pairwise No More: Rethinking Bias Detection Methods for Complex Intersectional Data
Sarah Quesen, WestEd, Brandon LeBeau, WestEd

Discussant:
Catherine Close

039. Innovations in the Assessment of Civic Learning in K-12 Schools
Coordinated Paper Session
9:45 AM - 11:15 AM
Hilton Denver City Center Lower Level 2: Colorado IJ

Civic learning, which refers to the broad set of competencies (skills, knowledge, and dispositions) needed for effective engagement in community and civic life, has received increasing attention in debates about the responsibilities of K-12 schools. Trends such as a changing media environment and increasing partisanship highlight the importance of effective and equitable civic-learning opportunities. High-quality assessment practices and tools are needed to monitor and improve those learning opportunities and to measure students' civic-related competencies as well as their civic engagement. Assessments to address the full set of opportunities and outcomes related to civics are limited, but advances in measurement methodology and other developments such as generative AI offer promising paths toward better, more useful assessments. In this session, authors from a forthcoming volume on assessment of civic learning opportunities and outcomes, to be published as part of the NCME Book Series, will discuss work related to advancing more equitable, innovative, and evidence-informed civics assessment. A discussant will synthesize themes and engage the audience in a conversation. The session's goals are to raise awareness of relevant work and help ensure that NCME members are equipped to advance this important application of assessment.

Session Organizer:
Laura Hamilton, American Institutes for Research

Chair:
Laura Hamilton, American Institutes for Research

Presentations:
How Generative and Predictive AI Can Power Innovation in Civics Assessment
Eric Tucker, Equity by Design; Eva Baker, University of California, Los Angeles

Can Discrete-choice Measures and Situational-judgment Tests Address Questions about Quantifying Global Citizenship?
Michael Thier, International Baccalaureate Organization, University of Oregon

Developing and Evaluating Civic Competencies through Scenario-Based Assessment
Alli Brettschneider, ETS; Greg Vafis, ETS; E. Caroline Wylie, Center for Assessment

Toward a Framework for Monitoring Equity in Civic Opportunities to Learn
Margarita Olivera Aguilar, American Institutes for Research; Samuel Rikoon, American Institutes for Research; Laura Hamilton, American Institutes for Research; Corey Savage, University of Glasgow

Assessing Student Engagement in a U.S. History and Civics Curriculum
Leah Bueso, University of Illinois Springfield; Marsha Ing, University of California, Riverside; Abigail Dym, Providence College; Joseph Kahne, University of California, Riverside

Discussant:
Doris Zahner, Council for Aid to Education

040. Innovations in Using Large Language Models for Automated Scoring and Automated Writing Evaluation

Individual Paper Session
9:45 AM - 11:15 AM
Hilton Denver City Center Lower Level 2: Denver 5-6

Chair:
Yue Huang, Measurement Incorporated

Presentations:

Automated Scoring of Student Essays for Diagnostic Assessments
Jason Bryer, City University of New York; Angela Lui, City University of New York; Antonio Fraser, City University of New York; Julia Ferris, City University of New York
Significant advancements in large language models have occurred over the past decade. This study explores how some of the more recent tokenizers compare to traditional n-gram-style tokenization procedures in the context of scoring a diagnostic assessment. Results show improvements, but predictive models might matter more.

Toward Subtrait-Level Model Explainability in Automated Writing Evaluation
Luis Alejandro Andrade-Lotero, Pearson; Lee Becker, Pearson; Joshua Southerland, Pearson; Scott Hellman, Pearson; Bradley Behan, Pearson
Subtrait (latent-trait components) assessment presents a promising path toward enhancing transparency of automated writing scores. We prototype explainability with generative language models and show high correlation between human subtrait and trait scores, and between automated and human subtrait scores. Our approach provides details to demystify scores for educators and students.

Integration of Machine Learning and Human Rater Scores using the Many-Facet Rasch Model
Xingyao Xiao, University of California, Berkeley; Richard Patz, University of California, Berkeley; Mark Wilson, University of California, Berkeley
This study extends the Many-Facet Rasch Model (MFRM; Lincare, 1989) to integrate machine learning and human rater scores for constructed response items. By employing an overlapping block design, we explore various rater assignment strategies, demonstrating the effectiveness of MFRM in enhancing reliability and validity in scoring, while managing rater variability.

Discussant:
Daniel McCaffrey, ETS

041. IRT Scoring and Rating Methods

Individual Paper Session
9:45 AM - 11:15 AM
Hilton Denver City Center Lower Level 2: Denver 3

Chair:
Annette Vernon, University of Iowa

Presentations:

Scoring SHARP Items with Testlet Response Model and its Variants
Bowen Wang, University of Florida; Richard Feinberg, NBME; Christopher Runyon, National Board of Medical Examiners
This study investigates the performance of three Rasch-based testlet models on a novel item type with small testlet size format. Preliminary findings suggest that the traditional Rasch model is robust when testlet effects are small and a Rasch-based fixed-effect testlet model performs best when testlet effects are large.

Using Mokken Scaling to Evaluate Rating Quality with Anchor Designs
Stefanie Wind, The University of Alabama; Mirai Nagasawa
Although nonparametric methods are useful for exploring rating quality, their typical application requires complete data. We consider how including anchor sets of examinees whose performances are scored by all raters to otherwise sparse rating designs can facilitate rating quality analyses under conditions with different anchor set characteristics and sample sizes.

Classification Accuracy and Consistency for Multiple Hurdles in the IRT framework
Seohee Park, American Board of Internal Medicine
This study presents methods for estimating classification consistency and accuracy in assessments with multiple hurdles, using ?-based cut scores. The simulation study covers all decision rules outlined by Chester (2003) and compares the proposed methods with the number correct cut score-based approach (Park et al., 2023).

Discussant:
Gozde Sirganci, Southern Methodist University

042. Leveraging Through-Year/Interim Assessment Results for Instructional Use

Individual Paper Session
9:45 AM - 11:15 AM
Hilton Denver City Center Lower Level 2: Colorado CD

Chair:
Rumeysa Gorgulu, Purdue University

Presentations:

Exploring Methods to Make Through-Year Assessment Results Instructionally Relevant
Pooya Razavi, Edmentum; Sonya Powers
We evaluate three methods for placing students into personalized learning programs using data from students who completed both a state assessment and an interim test. Using state assessment scores for student placement provides a mechanism for realizing the promise of through-year assessments to decrease testing burden and increase instructional relevance.

Exploring alternative IRT Models to improve estimation in the presence of learning
Jiawei Xiong, Curriculum Associates; Erin Banjanovic
Personalized instruction is presented in a different context from typical interim assessments. The current study explores the use of additional IRT model parameters to account for these differences and considers the impact on proficiency estimation. Preliminary results suggest discrimination and guessing parameters vary in informative ways that may aid estimation.

Patterns of use from an instructionally embedded assessment pilot study
Brooke Nash, University of Kansas; Jake Thompson; Auburn Jimenez, University of Kansas; Mary Majerus, Missouri Department of Elementary and Secondary Education; Shaun Bates, Missouri Department of Elementary and Secondary Education
Instructionally embedded and through-course assessments have garnered increased attention recently in a search for relevant and actionable assessment designs. In this paper, we describe teachers use patterns (e.g., content selection, student mastery) from an innovative assessment system pilot study to better understand how these systems are applied in practice.

Locating Optimal On-Track Cut Score in Through-Year-Assessment Designs using Receiver Operating Curves
Maryam Pezeshki, Cambium Assessment Inc.; Shuqin Tao, Cambium Assessment Inc.; Stephan Ahadi, Cambium Assessment Inc.
For through-year-assessment (TYA) designs, it is important to determine optimal cut scores for interim assessments to predict end-of-year proficiency. This study proposes the application of

Receiver Operating Characteristic (ROC) curves for identifying optimal cut scores for interim assessments using summative proficiency cut. Different methods are compared, and implications are discussed.

Discussant:
Bryan Drost, Rocky River City School District

043. Process Data Analyses to Understand Large-Scale Survey-Based Assessment Results
Individual Paper Session
9:45 AM - 11:15 AM
Hilton Denver City Center Lower Level 2: Denver 4

Chair:
Mingya Huang, University of Wisconsin -- Madison

Presentations:

Digital Literacy for Digital Assessments: Evidence from PIRLS 2021
Jeneve Swaby; Matthias von Davier, Boston College
This study investigates the relationships between digital self-efficacy, socioeconomic status (SES), reading confidence on students' time on task and reading performance in PIRLS 2021. Additionally, it explores how time spent reading relates to digital literacy and the role of nested data structures in influencing these relationships across schools and countries.

International Differences in Survey Characteristics Predicting Screen Response Time
Ella Anghel, Boston College; Matthias von Davier, Boston College
We tested whether the relationship between the number of items and response options per screen and response time varies internationally using the PIRLS student questionnaire. The number of items per screen was associated with response time in all countries, but the effect of the number of response options varied.

Discrepancy Between Self-Reported and Response Time-Based Questionnaire-Taking Efforts on Students' PISA Achievements
Surina He; Xiaoxiao Liu, University of Alberta; Ying Cui, University of Alberta
This study examines the impact of discrepancies between self-reported and response time-based questionnaire-taking efforts on performance, using PISA 2022 data from 20,560 Canadian students. Polynomial regression and response surface analysis indicated that students with congruent efforts outperform those with discrepancies, with higher RBE outperforming higher SRE when discrepancies occur.

Discussant:
Qiwei He, Georgetown University

045. Is Unaccounted-for Multidimensionality Really That Problematic?
Organized Discussion
9:45 AM - 11:15 AM
Hilton Denver City Center Lower Level 1: Penrose 2

Researchers have been interested in the multidimensional structures of latent constructs for over a century. There are myriad ways to model and evaluate these structures, and numerous instances where evidence of multidimensionality is identified; yet, in practice, very few assessments use multidimensional models for the purpose of item parameter estimation and scoring. Rather, constructs are typically treated as essentially unidimensional. For this session, four panelists have been assembled to debate theorical issues related to the identification and specification of multidimensional factor structures, issues related to unaccounted-for multidimensionality in applied contexts, and the extent to which multidimensionality is problematic for operational assessment.

Chair:
Jonathan Weeks, American Institutes for Research

Moderator:
Jonathan Weeks, American Institutes for Research

Presenters:
Jonathan Weeks, American Institutes for Research
Mark Reckase, Psychometric Solutions
Derek Briggs, University of Colorado - Boulder
Terry Ackerman, University of Iowa
Joseph Martineau, ETS

193. Innovation Demonstration 1
Electronic Board Session
9:45 AM – 11:15 AM
Hilton Denver City Center Lower Level 2: Colorado GH

Presentations:

1. **On-the-Fly Context Personalization of Mathematics Word Problems through Generative AI (Context-AI)**
   Burcu Arslan, ETS Research Institute; Candace Walkington, Southern Methodist University; Jesse Sparks, ETS; Matthew Bernacki, University of North Carolina at Chapel Hill
   Context-AI is a generative AI-powered prototype that personalizes math word problems in real time based on examinees' interests tied to their cultural identities. We will demonstrate the Context-AI's goals and theory of change, perform a live demo based on attendee input, and discuss the quality of the personalized items.

2. **An automated annotation model for science argumentation and reasoning in written assessments**
   Dante Cisterna, ETS; Lei Liu, ETS; Yi Song, ETS; Field Watts, ETS; Devon Kinsey, ETS
   The AI-SPIN Science project develops AI-supported assessments to support scientific reasoning and argumentation with special attention to English Language Learners' needs. We will demo an AI-tool that identifies reasoning features as students engage in argumentation and provides real-time feedback. Participants will interact with the assessments and discuss potential classroom applications.

3. **Automated Scoring of Mathematics Items Using Large Language Models (LLMs): An Automated Tool**
   Ruhan Circi, American Institutes for Research; Elana Margaret Perkoff, University of Colorado Boulder; Ozge Can Aran, Hacettepe University
   This presentation introduces an innovative scoring tool supported by GPT-4o to enhance the automated scoring of various mathematics items, including multiple-choice, short, and long constructed-response types. Using chain-of-thought prompting generated by LLMs, the system processes inputs to predict scores, maximizing transparency and automation in the steps.

4. **R Shiny Application for Multilevel Regression Discontinuity Analysis with Measurement Models**
   Muwon Kwon, University of Maryland; Youngjin Han, University of Maryland; Youjin Sung, University of Maryland; Ji Seung Yang, University of Maryland; Yang Liu, University of Maryland
   We introduce an R shiny application for multilevel regression discontinuity analysis with latent running and/or outcome variables measured by categorical and/or continuous responses. The web-based application allows researchers to examine heterogeneity and generalizability of average treatment effect by adopting the novel approach with a graphical interface and produce publication-ready figures.

5. **Increasing Educator Assessment Literacy: The Teacher Leader Assessment Specialist (TLAS) Program**
   Ellen Vorenkamp, Michigan Assessment Consortium; Tara Kintz, Michigan Assessment Consortium
   The TLAS program by Michigan Assessment Consortium is an innovative initiative, designed to address the widespread issue of assessment illiteracy among educators. This competency-based program empowers teachers to become Classroom Assessment Specialists (CAS), who then lead districts in the development and implementation of more coherent balanced assessment systems.

6. **Sentence-BERT: A Tool for Identifying Item Enemy Pairs During Test Form Assembly**
   Marc Kroopnick, Association of American Medical Colleges; Erica Rutzler, Association of American Medical Colleges; Ying Jin, Association of American Medical Colleges
   For many testing programs, form assembly represents one of the most difficult and time-consuming tasks required for the successful administration of the exam. This demonstration will show how an LLM (Sentence-BERT) was incorporated into the MCAT® exam's item bank to automate the identification of item enemy pairs during form assembly.

7. **A Flexible Automated Test Assembly Tool**
   Yanyan Fu, Graduate Management Admission Council
   A flexible ATA tool is built to allow users to assemble forms with a variable number of items that share a common stimulus, solve the maximum number of forms that can be built given content constraints, control overlapping items among the parallel forms, and prioritize items based on their usage.

8. **A Python-based Automated Test Assembly (ATA) System**
   Yi Gui, The University of Iowa; Jianshen Chen, College Board; Luz Bay, College Board
   A Python-based ATA system is developed using linear programming functions in the scipy package. Multiple uses will be demonstrated, including linear form assembly, item selection, and item pool split. The development process of the system will also be shared so that customized ATA systems may be built per users' needs.

9. **Score Reliability as Validity Evidence: Practical Considerations and a Shiny App**
   Kyle Nickodem, University of Minnesota - Twin Cities; Mohammed Abulela, University of Minnesota; Michael Rodriguez, University of Minnesota
   Score reliability is commonly misunderstood as an antecedent to validity rather than as a source of validity evidence. We demonstrate a Shiny app and user's guide that helps applied researchers and practitioners select the reliability coefficient most closely aligned with their intended score interpretation and use.

10. **A new approach for Psychometric Evaluation of MST's Theory and R functions**
    Cheng Hua, University of Montevallo; Louis Roussos, Cognia
    Utilizing Item Response Theory (IRT) simulations, our demonstration introduces a set of innovative R functions designed to calculate various Expected Test Information Function (ETIF) evaluation curves and indices. These functions, along with an R shiny app, empower researchers and practitioners to enhance multistage test quality and make informed data-driven decisions.

046. What has, can, and should happen when the Standards are not followed? (NCME Committee Organized Session)
Organized Discussion
9:45 AM - 11:15 AM
Hilton Denver City Center Lower Level 1: Penrose 1

As the Joint Standards for Educational and Psychological Testing undergo a review and update to reflect shifting current practices, advances in technology, and novel approaches to validity, questions remain about how the Standards will be applied in practice and used to defend or litigate assessment programs and processes. This timely panel discussion will explore how the Standards have been used in the past in law and policy and offer a dialogue around how they may be used in the future. The session will ask questions around what authority or "teeth" the Standards can and should have. Areas of discussion will include legal considerations for measurement professionals, non-legal enforcement mechanisms against violations of the Standards, and the equity lens in the revised Standards. This topic is relevant to all members of the NCME community, as well as to external stakeholders and partners in education.

Chair:
Regina Lewis, Maine Department of Education

Moderator:
E. Wyatt Gordon, Pearson

Presenters:
Stephen Sireci, University of Massachusetts Amherst
Jon Twing, HumRRO
Susan Davis-Becker, ACS Ventures, LLC

047. Social-Emotional Learning Measurements in Service of ALL Students: Lessons Learned around COVID
Coordinated Paper Session
11:30 AM-1:00 PM
Hilton Denver City Center Lower Level 1: Penrose 2

In a joint effort on the refinement and validation of a large-scale panel survey of students' social-emotional learning (SEL) in service of all students in grades 4 through 12, we bring together a diverse set of papers from researchers across institutions and organizations in this coordinated paper session. Working with students and longitudinal SEL survey data from the CORE Districts in California since the 2014-15 school year, we highlight lessons learned around survey measurement, the development, impact, and interpretations of student SEL measures, and differences in these findings between pre-COVID and COVID-impacted years. More specifically, researchers (1) discuss findings from cognitive interviews with students to answer questions "How do students in grades 4-12 conceptualize SEL items? How, if at all, do students' conceptualizations vary across English Language learner status, race/ethnicity, gender, and grade level?" (2) apply two-level bifactor, second-order factor, and correlated factor models to evaluate differentiations among the SEL constructs at both the school and student levels, (3) examine student longitudinal SEL developmental patterns, associations with academic outcomes, and differentiations among student subgroups using multilevel growth curve models, and (4) investigate the impact of the SEL constructs on students' chronic absenteeism status and academic outcomes.

Session Organizer:
Caroline Wang, Education Analytics

Chair:
Caroline Wang, Education Analytics

Presentations:

Improving Student Social-Emotional Learning Measurements in Service of ALL Students
Jose Mendez, UC Irvine; Kripa Sundar, EdTech Recharge; Jessica Hoskins; Andrea Macias, University of Southern California; Jasmin Hernandez, UC Irvine; Guadalupe Mendoza, UC Irvine; Katerina Schenke, ETS

Understanding Distinctions in Student-School Social-Emotional Measures Using Multilevel Factor Models
Daniel Bolt, University of Wisconsin – Madison; Caroline Wang, Education Analytics; Robert Meyer, Education Analytics & University of Wisconsin at Madison

The Developmental Curve of Students' Social-Emotional Competencies and COVID's Impact
Caroline Wang, Education Analytics; Shuyuan Li, Education Analytics

The Impact of Students' Social-Emotional Learning on Chronic Absenteeism and Students' Achievement
Sara Hu, Education Analytics; Robert Meyer, Education Analytics & University of Wisconsin at Madison

Discussant:
Michael Rodriguez, University of Minnesota

048. Fuse Diverse Process Data in Educational Measurement

Coordinated Paper Session
11:30 AM - 1:00 PM
Hilton Denver City Center Lower Level 1: Mattie Silks

The increasing popularity of digital testing environments affords the collection of a wealth of data recorded in log files, including response time, test-takers' actions, and textual data from constructed responses. Despite the richness of this data, traditional research methods often do not adequately capture and incorporate additional information. Data fusion-related methodologies offer a promising solution by integrating process data with traditional response data, creating a more dynamic and comprehensive picture of test-taking behaviors and educational outcomes.

This session presents innovative methods that leverage process data to address testing and education problems. Specifically, we will discuss how modeling test-takers' responses and response times as sequential data using Long short-term memory enhances adaptive testing; how longitudinal latent space modeling identifies time-varying patterns in test-takers behaviors; how a bi-level, multi-dimensional scoring model for constructed responses improves writing analysis; how machine learning models identify key variables that explain achievement gaps; and how integrating textual features from constructed responses into IRT ability estimation improves performance predictions in mixed-format assessments.

Session Organizer:
Huan Kuang, Florida State University

Chair:
Susu Zhang, University of Illinois at Urbana-Champaign

Presentations:

Sequential modeling of responses and response times in a CAT with LSTM
Joyce Xinle Liu, University of Alberta; Okan Bulut, University of Alberta

Longitudinal latent space modeling of process data
Xin Qiao, University of South Florida; Tracy Sweet, University of Maryland

Neural topic model for interpretable multi-dimensional automated scoring of constructed responses
Cheng Tang, The University of Georgia; Qidi Liu, GlobalFoundries; Feiming Li, Zhejiang Normal University; Jiawei Xiong, Curriculum Associates

Exploring performance differences based on process data variables using machine learning techniques
Bowen Wang, University of Florida; Huan Kuang, Florida State University; Jiawei Xiong, Curriculum Associates

Mixed-format assessment scoring: integrating textual features from topic models in IRT ability estimation
Jordan Wheeler, University of Nebraska – Lincoln; Shiyu Wang, University of Georgia; Allan Cohen, University of Georgia

Discussant:
Hong Jiao, University of Maryland

049. Exploring Methods to Expand Reporting on NAEP for Groups of Students

Coordinated Paper Session
11:30 AM - 1:00 PM
Hilton Denver City Center Lower Level 2: Denver 5-6

Labeled "The Nation's Report Card", the National Assessment of Educational Progress (NAEP) offers the only common metrics of achievement on which representative samples of students from all 50 states and several large school districts can be compared. Importantly, NAEP also disaggregates results by subgroups and types of students, which enables policymakers to develop a sense of how different groups of students are achieving across jurisdictions.

Unfortunately, achievement estimates for many groups of students are omitted from NAEP reporting. To prevent the use of imprecise estimates, NAEP only reports estimates calculated from samples of at least 62 students (Elliott & Phillips, 2004). In other cases, to prevent the use of estimates potentially impacted by nonresponse bias, estimates are omitted when an insufficient percent of sampled schools participate in the assessment (National Center for Education Statistics, n.d.).

These omissions present a limitation for policymakers interested in understanding how different groups of students (e.g., Black students, private school students), including underserved groups, are achieving across jurisdictions. This coordinated paper session presents four research studies using NAEP data that explore the use of statistical methods to overcome common sampling limitations—research that may someday support expanded reporting for groups of students on NAEP.

Session Organizer:
David Bamat, American Institutes for Research

Presentations:
Leveraging state assessment data to improve a common Small Area Estimation model
David Bamat, American Institutes for Research

An Innovative Way of Estimating Subgroups Mean Estimates: SABDB
Sinan Yavuz, American Institutes for Research

Expanding NAEP Subgroup Reporting with Random Forests
Paulina Grekov, University of Wisconsin-Madison; David Bamat; Sinan Yavuz, American Institutes for Research

Discussant:
Gary Phillips, Cambium Assessment

050. On the Nature and Rigor of Classroom Assessment

Coordinated Paper Session
11:30 AM - 1:00 PM
Hilton Denver City Center Lower Level 2: Denver 1-2

Questions about the relationship between classroom assessment and educational measurement, and especially whether classroom assessment is a kind of educational measurement or a separate, though related, field, have been under discussion for some time (Brookhart & McMillan, 2020; Ho et al., 2024; McMillan, 2013). The idea that classroom assessment is a separate field is based on the argument that it does not meet the standards of measurement: classroom assessment does not need high rigor or attention to reliability, because inappropriate decisions made in the classroom can easily be corrected during the flow of instruction with no harm done. The purpose of this symposium is to explore the nature and rigor of classroom assessment. Two papers will explore claims about the relationship between classroom assessment and educational measurement, specifically whether and under what conditions classroom assessment is a form of educational measurement. Two papers will explore what it means to be rigorous in classroom assessment, in the sense of being both thorough and precise, in dispute of the argument that rigor in classroom assessment matters to a lesser degree than it does in large-scale assessment.

Session Organizer:
Susan Brookhart, Duquesne University

Chairs:
Susan Brookhart, Duquesne University
Sarah Bonner, Hunter Collete, CUNY

Presentations:
When is Classroom Assessment Educational Measurement?
Susan Brookhart, Duquesne University; Sarah Bonner, Hunter College, CUNY

What Do NCME Members Need to Understand about Classroom Assessment?
Dustin Van Orman, Western Washington University; Heidi Andrade, University of Albany

Validity in the Context of Classroom Assessment
Maria Ruiz-Primo, Stanford University; Guillermo Solano-Flores, Stanford University

Formative Assessment and Educational Measurement
Randy Bennett, ETS

Discussant:
James McMillan, Virginia Commonwealth University

051. Graduate Student eBoard Session 1

Graduate Student Electronic Board Session
11:30 AM - 1:00 PM
Hilton Denver City Center Lower Level 2: Colorado GH

Session Organizer:
Rumeysa Gorgulu, Purdue University

Presentations:
1. **Game-Based and Gamification-Based Assessments' Effectiveness on Reducing Test Anxiety: Systematic Review**
   Rumeysa Gorgulu, Purdue University; Yukiko Maeda, Purdue University
   This systematic review synthesizes empirical evidence reported from 2011 to evaluate the effectiveness of game-based and gamification-based assessments (GBAs) in reducing test anxiety.

A total of 21 studies across various educational levels were analyzed to explore the impact of game elements in assessments on alleviating test anxiety and students' well-being.

2. **Measuring Oral Reading Skills Based on Silent Reading Measures**
   Tonya Murray, Stanford Graduate School of Education; Wanjing (Anya) Ma; Jasmine Tran, University of California, Irvine; Albu Ungashe, Stanford University; Kelly Wentzlof, Stanford University; Mia Fuentes-Jimenez; Carrie Townley-Flores, Graduate School of Education, Stanford University; Jason Yeatman, Stanford University
   Individually administering assessments that feature oral reading of individual words and passages is time-consuming and resource-intensive, and may be affected by administrator bias. This paper demonstrates that the silent reading measures implemented in ROAR-Word and ROAR-Sentence and administered to whole classrooms correlate strongly with traditional, orally administered measures.

3. **What Do We Know About Innovative Assessment?**
   Xiaohang Zhang
   It's hard to read about assessment these days without coming across the term, innovative assessment. In this paper, I explore the concept of innovation by drawing on work from many fields. I then use this conceptualization to create criteria that I apply to existing assessments and potential innovative assessments.

4. **Evaluation of Distractor Models within Assessment Engineering Using the Think-Aloud Method**
   Sunhyoung Lee, Ascend Learning/ University of Nebraska-Lincoln
   This study evaluates distractor models within the Assessment Engineering framework, focusing on their alignment with construct knowledge and their impact on students' problem-solving. By integrating psychometric properties with the think-aloud method, it aims to provide practical insights for improving the quality of multiple-choice questions and guidelines for effective distractor model development.

5. **Measuring Student Agency Conditions**
   Nicholas Balisciano, Harvard University
   A literature review of definitions and measures of student agency against the tenets of Social Cognitive Theory revealed fundamental gaps. To help address this, I created the Student Agency Conditions Scale. Its ongoing development has been informed by critical and institutional sociology and involves both student and educator collaboration.

6. **Using preference and discrimination indices for block constructing in forced choice tests**
   Sirui Wu, University of British Columbia
   This study proposed a item selection method for multidimensional forced choice test, optimizing both preference and discrimination indices. It leads to improved accuracy and precision compared to random pairing, as evidenced by lower RMSE and higher test-criterion relationship correlations using a simulation study.

7. **Co-Design(-ing) to Align Reading Assessments with Workplace Activities in the Agricultural Industry**
   Aria Immanuel, University of Massachusetts Amherst; Maria Oliveri, Purdue University
   In this study, we used a co-design approach to develop workplace-aligned, authentic reading language assessments in the agricultural industry based on combining results from job analyses, interviews, and observational data with national survey data. We also explored methods to broaden the generalizability of the assessment approach to other occupations.

8. **Automated Test Assembly with R Packages: Comparison of ROI, ompr and eatATA**
   Huijuan Wang; Keith Boughton, Data Recognition Corporation; Christine DiStefano, University of South Carolina

This study compares three R packages ROI, ompr, and eatATA for Automated Test Assembly (ATA) through mixed-integer programming (MIP) with two practical scenarios. Results yielded similar outcomes across packages, with eatATA being user-friendly for beginners, while ROI and ompr offer greater flexibility for advanced users and more complex ATA tasks.

9. **Work in progress: Identifying misconceptions in successive assessment in STEM education.**
   Amirreza Mehrabi, Purdue University; Jason Morphew, Purdue University
   Understanding student misconceptions is vital for enhancing learning in engineering education. We propose a Transition Diagnostic Classification Model that combines CDM and transitional models to identify misconceptions as attributes and provide targeted feedback, preventing misconception persistence and development, thereby improving assessment accuracy and addressing learning gaps effectively.

10. **Meta-Analysis of the Impact of Standardized Testing on Educational Equity: Societal Perspective**
    Clement Yeboah, The University of Southern Mississippi
    This meta-analysis of 15 studies (45,678 participants) found that standardized testing slightly increases educational inequity, particularly among marginalized groups, with a summary effect size of 0.32. Despite aiming for uniform measurement, standardized tests may unintentionally perpetuate social disparities, highlighting the need for more equitable assessment practices.

11. **Exploring Classroom Heterogeneity Based on Students' English Language Proficiency Levels**
    Roza Nalbandyan
    Most research on tracking focuses on race and class, not English Language classification. We examine how mathematics class heterogeneity differs by language classification using 2,500 6-8 grade student data. Network methods reveal course-taking by language designations disparities. Markov's model reveals prior enrollment increases the likelihood of certain mathematics trajectories.

12. **Exploring eight PROMIS® profile domain scores and an overall health score with two MIRT models**
    Lihua Yao, Northwestern University; Anne Huang, Scarsdale High School, New York; XiaoDan Tang, Northwestern University Feinberg School of Medicine Department of Medical Social Sciences
    Two multidimensional IRT (MIRT) models were applied to the PROPs scores: an eight-dimensional MIRT and a higher-order MIRT. The scores from each MIRT model were compared to those from the unidimensional model. The goal was to explore the possibility of reducing the number of assessment items without sacrificing prediction accuracy.

13. **Assessing Bayesian Methods for Longitudinal Drift Detection in Small Samples with Non-Normal Data: A Simulation Study**
    Olushola Soyoye, University of Delaware; Kamal Chawla, University of Maine; Jesse Eze, University of Delaware
    Drift in item performance threatens assessment validity. This study evaluates Bayesian methods for detecting item parameter drift (IPD) in small samples with non-normally distributed data, comparing them to traditional frequentist approaches. The focus is on the accuracy and robustness of these methods, with implications for improving assessment validity in psychometrics.

14. **Examining Sample Size Requirements for the Multidimensional Graded Response Model with Different Response Categories**
    Aaron Beuoy, University of Kansas
    The current simulation study examines sample size requirements for the multidimensional graded response models (MGRM) with varying numbers of response categories, dimensions, items per dimensions, and dimensional correlations. Implications and practical recommendations for the noncognitive test construction and evaluation using the MGRM are provided.

15. Advancing Sequence Clustering with Time-Wrapping Distance Measurement
Binhui Chen; Qiwei He, Georgetown University
A novel approach is proposed in this study to leverage Dynamic time Warping (DTW) and the Time Warped Longest Common Subsequence (T-WLCS) method, with the aim to better identify the behavioral patterns by clustering time-interval adjusted action sequences. A problem-solving task in PIAAC is used for an illustration.

16. Investigating Use and Enhancements of the NAEP Racial/Ethnic Score Gap Tool
Nicholas Copeland, University of Washington; Katherine Furgol Castellano, ETS; Marisol Kevelson, ETS
The NAEP racial/ethnic score gap tool, while successful in its effective presentation of psychometrically reliable data, may contribute to deficit thinking both explicitly and implicitly. This study focuses on developing and testing several alternative, asset-based and further contextualizing data visualizations in R-Shiny and proposes them as a viable alternative score reporting tool.

17. Social Support, Emotional Self-Efficacy: Impact on Academic Stress and Test Anxiety
Justice Dadzie, The University of Alabama; Ifesinachi Ezugwu, The University of Alabama; Stacy Surman, The University of Alabama
This study examines the relationships between academic stress, test anxiety, regulatory emotional self-efficacy (RESE), and social support among 370 students in the Sekondi-Takoradi Metropolis. Results indicate that RESE and social support mitigate test anxiety, suggesting the importance of emotional regulation skills and robust support systems.

18. Investigating the Impact of K-12 Enrollment Trends on Interpretations of Longitudinal Achievement
Mara McFadden, James Madison University
Enrollment declines post-pandemic have been a topic of concern among education researchers. This study examines school-level enrollment shifts and their impact on interpretations of performance trends over time. Results indicate that schools experiencing pronounced enrollment change should consider the impact on their interpretations of longitudinal achievement.

052. Improving Scoring for CAT

Individual Paper Session
11:30 AM - 1:00 PM
Hilton Denver City Center Lower Level 2: Denver 3

Chair:
Kristin Morrison

Presentations:

The Impact of Scoring Polytomous Items Later Using the Shadow Test Approach
Jing Ma, University of Iowa; Anthony Fina, University of Iowa; Stephen Dunbar, University of Iowa
This study investigated the impact of scoring polytomous items later on measurement precision and test security in mixed-format adaptive testing. Simulations using the shadow test approach across various designs showed that while scoring later had a statistically significant effect, its practical impact on outcomes was minimal.

Estimation of Conditional Standard Errors of Measurement for MLE Scores in MST
Yuanyuan Stirn; Won-Chan Lee, University of Iowa
This paper examines the accuracy of the conditional standard error of measurement calculated using an information-based analytic method for maximum likelihood estimates in multistage testing. As the test length increases, the analytic results become more similar to those obtained from the simulation approach.

Repercussions of Switching from MAP to MLE Scoring in Computerized Adaptive Testing
Adam Wyse, Renaissance; Edison Choe, Renaissance
This study uses a simulation and a real data example to show that when mixed response strings are obtained and CATs switch from MAP to MLE for interim ability estimation, the target difficulty of the next item and the average percentage correct on it can exhibit unexpected patterns.

Allowing Answer Changes in Computerized Adaptive Tests: Yes, We Can!
Jaime Malatesta, GMAC; Kyung (Chris) Han, GMAC
This study suggests a few options for allowing answer changes in Computerized Adaptive Tests (CATs) and uses a simulation and real data analysis to demonstrate how measurement efficiency and accuracy can be managed while allowing for answer changes in the CAT-based GMAT for enhancing the test taking experience.

Discussant:
Jonathan Beard, College Board

053. Measurement Invariance and Differential Item Functioning

Individual Paper Session
11:30 AM - 1:00 PM
Hilton Denver City Center Lower Level 2: Colorado CD

Chair:
Lucy Gitiria, UMass Amherst

Presentations:

Exploring Fairness of Next-Generation Items: Differential Item Action Analysis on Process Data
Yusuf Kara, University of Miami; Gozde Sirganci, Southern Methodist University; Xin Qiao, University of South Florida; Akihito Kamata; Sarunya Somsong, Srinakharinwirot University & Southern Methodist University; Michael Fong, Southern Methodist University
This study explores differential item actions in complex problem-solving tasks in digitally-based assessments. We explore the utility of mixture item response theory models applied to refined process data that reflect action sequences of responders. The results of the preliminary analyses provide implications related to the fairness of such complex tasks.

Detecting Differential Item Functioning in Process Data with Sequential Modeling
Yuting Han, Beijing Language and Culture University; Hongyun Liu, Beijing Normal University; Feng Ji, University of Toronto; Yunxiao Chen, London School of Economics and Political Science; Kaiyu Gan, Beijing Normal University
This study introduces the Sequential Response Model with Covariates (SRM-C) to analyze differential item functioning (DIF) in process data from computer-based assessments (CBAs). The SRM-C combines measurement and structural components to model response sequences while accounting for covariates and group-specific ability distributions. Simulation and empirical results demonstrate the model's effectiveness.

Understanding Measurement Utility of Technology-Enhanced Items
Cassondra Griger, The University of Iowa; Louis Roussos, Cognia; Qi Qin, Cognia
This study investigates the differences in measurement precision between simulated data sets of technology-enhanced items (TEIs) and traditional, paper-based items. Using differential beta functioning, we compare polytomous items to identify which item features provides more accurate assessment across varying ability levels.

Strengthening Cross-Grade Comparisons on a Vertical Scale with Regularized DIF Analysis
Sanford Student, University of Delaware
This study investigates the application of regularized moderated nonlinear factor analysis to identify and mitigate biasing effects of differential item functioning in vertical scaling. While unmodeled DIF can substantially bias traditional IRT-based vertical linking, the proposed method appears to substantially mitigate this bias when DIF is present.

Discussant:
Kevin Krost

054. Rapid Guessing and Aberrant Responses

Individual Paper Session
11:30 AM - 1:00 PM
Hilton Denver City Center Lower Level 2: Denver 4

Chair:
Jiawei Xiong, Curriculum Associates

Presentations:

Understanding the Behavior and Implications of Rapid Guessing
Dihao Leng, Boston College; Yanyan Fu, Graduate Management Admission Council; Kyung (Chris) Han, GMAC
Limited empirical research exists on examinees' guessing behavior. This empirical study found that rapid guessers—instead of guessing at random, as commonly assumed in achievement estimation—exhibited a systematic "edge aversion" tendency, particularly favoring "C" and avoiding "E". This pattern is consistent across various demographics. Psychometric implications are investigated.

Modeling Inattentive Responding Behaviors and Extreme Response Style in Time-based Likert-Type Data
Jimmy de la Torre, The University of Hong Kong; Zechu Feng, The University of Hong Kong; Kuan-Yu Jin, Hong Kong Examinations and Assessment Authority
Likert-type scales are commonly employed in educational and psychological surveys. However, inattentive responding behaviors can result in degradation of the construct estimates. This paper proposes a new item response theory model for distinguishing between attentive and inattentive respondents, and account for extreme response style (ERS) of the attentive respondents.

How Rapid Guessing Biases Parameter Estimates and Compromises Common Assessment Practices
Sarah Alahmadi, ACS Ventures, LLC; Christin Demars, James Madison University
Using bootstrapping methods to mimic data observed in testing contexts susceptible to low examinee effort, we clarified ambiguity in current literature by identifying conditions where item parameter estimates are likely to be overestimated or underestimated when rapid guessing is present, revealing serious implications for common assessment practices.

A Mixture IRtree Model for Aberrant Response and Missing Data
fangbin chen; Jiyuan Ding, Dongbo Tu, Jiangxi Normal University; Yan Cai, Jiangxi Normal University; Daxun Wang, Jiangxi Normal University; Junhuan Wei, Jiangxi Normal University
In tests, examinees may engage in behaviours such as rapid-guessing, cheating and non-response. These behaviours can result in inaccurate estimates of examinees' ability, thus undermining the validity and fairness of the test. So this paper proposes a mixture IRTree model that simultaneously considers normal response, rapid-guessing, cheating and non-response behaviours.

Discussant:
Hongwen Guo, ETS

055. Implications of Culturally Responsive Assessment for Large-Scale Assessment Practices (NCME Edited Volume)

Invited Session
11:30 AM - 1:00 PM
Hilton Denver City Center Lower Level 2: Colorado IJ

This session describes how a culturally responsive lens can shape assessment purpose, design, implementation, use, and score validation. The panelists will introduce a NCME edited volume that explores the conceptual foundations and implications of culturally responsive assessment in K-12 large-scale and classroom contexts. The goals for the session are to provide current perspectives on culturally responsive assessment within large-scale assessment and to prompt discussion among the panelists and the audience about how measurement and assessment must change to better align with a pluralistic and multicultural society.

The editors will provide an overview of the book, its purpose, and organization. Authors of several chapters will present the main themes from their work. We will then open the session to provide participants with an opportunity to ask questions and discuss the ideas from the chapters.

Session Organizer:
Carla Evans, Center for Assessment

Chair:
Carla Evans, Center for Assessment

Presentations:

A Framework for Enacting Equity Aims in Assessment Use: A Justice-Oriented Approach
Susan Lyons, Lyons Assessment Consulting; Maria Oliveri, Purdue University; Mya Poe, Northeastern University

A Measurement Argument for Culturally Responsive Assessment
Catherine Taylor; Steve Ferrara, HumRRO

Anti-Racist Approaches to Scoring Large-Scale Assessments
Lauren White; Jaylin Nesbitt; Heather Roeters-Solano, Pearson; Sarah Quesen, WestEd; Susan Lottridge, Cambium Assessment; Karen Lochbaum, Pearson

On Objectivity, Multicultural Education, and Large-Scale Assessments—Commentary on Section 2
Wayne Au, University of Washington Bothell

056. Test Security in the AI Age: Trends, Challenges, and Visions

Organized Discussion
11:30 AM - 1:00 PM
Hilton Denver City Center Lower Level 1: Penrose 1

This discussion session brings together three renowned security researchers to explore the evolving landscape of test security in the age of artificial intelligence. The speakers will share insights into current trends, emerging challenges, and innovative solutions for detecting aberrant test-taking behavior and strengthening test security. Their presentations will be followed by an interactive dialogue with the audience, fostering cross-organizational collaboration and identifying key areas for future research. By addressing the complexities introduced by AI, this session aims to contribute to the development of robust test security frameworks capable of confronting modern threats.

Chair:
Huijuan Meng, Amazon Web Services

Moderator:
William Belzak, Duolingo

Presenters:
James Wollack, University of Wisconsin – Madison
Kirk Becker, Pearson VUE
Duanli Yan, ETS

057. Honoring the Contributions of Steve Ferrara: Cognitive Complexity

Organized Discussion
11:30 AM - 1:00 PM
Hilton Denver City Center Lower Level 2: Colorado EF

In August, 2024, the field of educational measurement lost a leader in innovation, a valued colleague, and a dear friend: Steve Ferrara. One of Steve's favorite and most enduring areas of inquiry was the relationships among item design, empirical item difficulty, and cognitive complexity. Through his work in this area, he collaborated with and mentored many. In this session, several of his colleagues and mentees will gather in his honor to remember Steve and his contributions. Each panelist will be asked to share a story about Steve and how he influenced our work and thinking in the area of cognitive complexity. We will also speak to what kind of work lies ahead in this line of inquiry that the field needs to continue to innovate and improve, and that Steve would want us to pioneer to honor his legacy and keep one of his research passions alive. Ample time will be left for Steve's colleagues and friends in the audience to also share their memories and his contributions to their work and thinking.

Chair:
Kristen Huff, Curriculum Associates

Moderator:
Kristen Huff, Curriculum Associates

Presenters:
Kristen Huff, Curriculum Associates
Christina Schneider
Ellen Forte, edCount, LLC
Daniel Lewis, Creative Measurement Solutions
Jeffrey Steedle

058. Applications of Multistage Testing in the Age of Digital and AI Technology

Coordinated Paper Session
1:30 PM - 3:00 PM
Hilton Denver City Center Lower Level 2: Denver 1-2

In the age of digital and AI technology, information from all directions is overwhelming. We have seen significant progress in artificial intelligence (AI) and machine learning (ML), big data management, and increased computational power and in the proliferation of their applications across industries. This is also true in the applications of adaptive and multistage testing, a form of AI, which has been leading AI research since the 1960s, such as the algorithms for merging adaptive learning and adaptive testing.

This symposium focuses on the recent applications of adaptive multistage testing. It illustrates the latest practical applications with design considerations, methodological approaches, NLP, large scale operational implementation such as the 2024 Scholastic Aptitude Test (SAT), and new development of R packages (catR and mstR). The presentations are authored by renowned psychometricians recognized globally for their expertise in adaptive and multistage testing.

Session Organizer:
Duanli Yan, ETS

Chair:
Duanli Yan, ETS

Presentations:
Overview of Adaptive Multistage Testing in the Digital and AI Age
Duanli Yan, ETS; David Weiss, University of Minnesota

Module Assembly and Routing of Cognitive Diagnostic Multistage Adaptive Test
Manqian Liao, Duolingo; Hong Jiao, University of Maryland

Improving Literacy by Integrating Advances in the Learning Sciences and Natural Language Processing in Multistage Testing
Paul Deane, ETS; Tenaha O'Reilly; Duanli Yan, ETS; Zuowei Wang, ETS; Jonathan Weeks, American Institutes for Research

Considerations in the MST Design for the New Digital SAT Suite of Assessments
Thomas Proctor; Oliver Zhang, College Board

New Development of Computerized Adaptive and Multistage Testing Software
Ye Ma, Amazon Web Services; Duanli Yan, ETS

Discussant:
Alina von Davier, Duolingo

059. Gaining New Insights Through Analysis of Multimodal, Multi-Source Data

Coordinated Paper Session
1:30 PM - 3:00 PM
Hilton Denver City Center Lower Level 2: Denver 3

This proposal investigates the benefits of integrating multimodal, multi-source data to gain deeper insights into educational phenomena. Paper 1 examines the use of Latent Dirichlet Allocation and large language models to categorize key topics from text corpora derived from interview transcripts. Qualitative analysis of these transcripts will be used to develop future interview items. Paper 2 validates self-reported engagement by comparing them with behavioral indicators of engagement and self-reported and behavioral indicators of procrastination, using a multi-trait, multi-method framework and a zero-inflated model, and predicts student learning outcomes based on engagement. Paper 3 proposes the MultVizAI platform to explore multimodal large language model approaches for creating a comprehensive framework that assesses students' ability to visualize data through multimodal data, including text descriptions, datasets, and visual images. Paper 4 investigates the censored lognormal distribution and parameter estimation affecting test-takers' response times. Paper 5 introduces a novel method combining item responses, response time, and biometric data. This approach addresses the gap in utilizing eye-tracking data and enhances the traditional methods, improving our understanding of student behavior and strengthening exam security in education. Collectively, these papers contribute to advancing educational research by integrating diverse data sources, thereby offering more robust and actionable.

Session Organizer:
Anh Thu Le, University of Notre Dame

Chair:
Ying Cheng, University of Notre Dame

Presentations:

Integrating LDA Topic Modeling and LLM Prompt Engineering to Analyze Transcripts
Teresa Ober, ETS; Karyssa Courey, Rice University; Michael Flor, ETS

Assessing Engagement in Academic Contexts: A Multi-Method Validation
Anh Thu Le, University of Notre Dame; Ying Cheng, University of Notre Dame

MultVizAI: Enhancing Data Visualization Feedback with Multimodal Large Language Models
Bo Pei, University of South Florida; Xiao Li, University of South Florida

A Censored Distribution for Response Time
Qizhou Duan; Ying Cheng, University of Notre Dame

Assessing Testing Behaviors: Joint-Modeling with Multimodal Response Data and Visual Patterns
Kaiwen Man, University of Alabama; Yaohui Liu, University of Alabama; Qipeng Chen, University of Alabama

Discussant:
Qiwei He, Georgetown University

060. Automating Item Development with Artificial Intelligence: From Generation to Evaluation

Coordinated Paper Session
1:30 PM - 3:00 PM
Hilton Denver City Center Lower Level 2: Colorado AB

Developing assessment items is a complex process spanning numerous phases including item generation, alignment, calibration, deployment, analysis, and evaluation. This process relies on extensive human expertise and effort, resulting in a prolonged life cycle for item development. Large language models (LLMs) have been applied to automate and accelerate item generation in a manner that has often matched or even exceeded human performance. While most existing works in using AI for item development, naturally, focus on using LLMs for automated item generation, there is an urgent need to carefully investigate the impact of LLMs on other phases of the item development cycle, especially calibration and evaluation. This session strikes a balance between item generation and evaluation at all phases of item development: three papers focus on novel approaches to item generation, showing that LLM-generated items can perform exceptionally well and often match the performance of humans. The other three papers focus on item evaluation, introducing novel evaluation methods and metrics that reflect real-world item use cases and item design principles, achieving high alignment with the judgement of human experts. The presenters consist of junior and senior researchers from nonprofit organizations, assessment companies, and universities.

Session Organizer:
Andrew Lan, University of Massachusetts, Amherst

Chair:
Andrew Lan, University of Massachusetts, Amherst

Presentations:

Automatic Chart Item Generation for Medical Assessment with Large Language Models: Challenges of Accuracy, Novelty, and Evaluation
Victoria Yaneva, National Board of Medical Examiners

Aligning Automation with Expertise: A Better Approach to Assessment Evaluation
Steven Moore, Carnegie Mellon University

Automatic Question Generation for Formative Practice: Setting the Benchmarks for Performance Metrics using Student Data
Rachel van Campenhout, VitalSource

Investigating Different Prompt Engineering Strategies Across Different Content Areas
Sergio Araneda, Caveon

Multiple Choice Question Generation: LLMs can Match Human Performance
Andrew Olney, University of Memphis

AI Generation of Medical specialty Items with LLM
Henry Makinde; Marcus Walker; Andrew Dallas

Discussant:
Christopher Ormerod

061. Graduate Student eBoard Session 2

Graduate Student Electronic Board Session
1:30 PM - 3:00 PM
Hilton Denver City Center Lower Level 2: Colorado GH

Session Organizer:
Peter Tran, University of Massachusetts, Amherst

Presentations:

1. **Evaluating the Content Quality of Items Written With and Without Artificial Intelligence**
   Peter Tran, University of Massachusetts, Amherst; Ketan, University of Massachusetts, Amherst; Stephen Sireci, University of Massachusetts Amherst
   Generative artificial intelligence (AI) is increasingly being used to develop educational test items without much research on their quality. This study compares alignment ratings across items written by humans and "cloned" versions developed using AI. The results will show how using AI in item development affects content validity.

2. **Examining the Effects of AIG Item Variance on Equating Outcomes**
   Aurore Phenow
   What happens when AIG item siblings are used as anchor items for equating? This study used a simulation to study the effects of within-family variation of item siblings on equated item and person parameters and linking coefficients. The findings offer insights into their practical implications for equating.

3. **School Climate and Student Mathematics Achievement: A MSEM approach**
   Hyunjung Lee, Fordham University; Chansoon (Danielle) Lee; Heining Cham, Fordham University
   This study examines the effects of school climate-related factors on 8th grade students' mathematics achievement using multilevel structural equation modeling. The findings of this study point out the importance of enhancing school safety, allocating more resources to mathematics, and improving classroom climates to increase students' mathematics achievement.

4. **Analyzing School-level K-3 Reading Trends Post COVID-19 with Quantile Panel Regression Estimators**
   Havisha Khurana, University of Oregon; Patrick Kennedy, University of Oregon; Gina Biancarosa, University of Oregon
   The COVID-19 pandemic caused widespread declines in student test scores. This study uses quantile regression growth models to examine how reading scores across the distribution changed from 2019-20 to 2023-24. Using a widely used progress-monitoring assessment, we address clustering challenges in quantile regression and compare multiple estimators to capture shifts.

5. An IRTree Approach to Investigating Item-level Predictors of Disengagement on the PISA.
Juste Mehou, James Madison University; Brian Leventhal, James Madison University
Disengagement in low-stakes assessments, like PISA, undermines scores interpretation validity. Identifying item characteristics that contribute to disengagement is crucial for effective test design, which could lower disengagement levels and improve the validity issue. This work uses an IRTree model to estimate ability and disengagement simultaneously and predict its item correlates.

6. Handling Missing Responses: Comparing the Effectiveness of Different Methods in Ability Estimation
Tram-Anh Tran Nguyen, University of Massachusetts, Amherst; Justin Kern, University of Illinois at Urbana-Champaign
Missing responses are common in educational assessments. This study compares the effectiveness of three approaches - IRTree, multiple imputation, and treating missing responses as not administered - in estimating test-takers' proficiency. While all methods produced low bias, the IRTree method was the most effective, consistently yielding low bias and RMSE.

7. Effects of Various Multidimensional Testlet Models on Ability Standard Errors
Brian Harrold, University of South Carolina; Brian Habing, University of South Carolina; Louis Roussos Cognia
Unidimensionality and local independence assumptions underly the analysis of many large-scale assessments. The testlet (or bifactor) structure violates these and is presumed to result in an underestimation of the variability in ability estimates. This study will investigate similar appearing exam structures that maintain their reliability of estimates.

8. Simultaneous linking on independent group calibration in TIMSS Background scaling
Tianzheng Mao, Boston College; Matthias von Davier, Boston College
Traditional moment-based and concurrent-calibration linking methods face potential bias and errors when linking independent calibrations across multilingual groups in Large-Scale Assessment background-scaling. Simultaneous linking (Haberman, 2009) transforms item parameters into a comparable scale while minimizing transformation error. This research tests its feasibility and compares its performance to the alignment methods.

9. A Latent Profile Analysis of Students' Classroom Environments Across COVID-19
Jay Jeffries, University of Nebraska-Lincoln
This study found six emergent profiles of student classroom environments. One profile, characterized by low concerns for bullying and peer conflict, was significantly less prevalent amidst the COVID-19 pandemic. These findings underscore the importance of regular classroom screenings to identify students in distress and provide timely support during crises.

10. The Growing Importance of Social Skills? A Measurement-Oriented Reanalysis of Deming (2017)
Nora Delaney
Socioemotional skills are widely agreed to be important for children's educational development, despite disagreement over their precise definitions and measurement. I reanalyze the results of Deming (2017), showing that a rigorous measurement approach refines an understanding of the relative importance of social skills in the K-12 context and beyond.

11. Validation of a Measure of Teachers' Math Knowledge
Joanne Joo, Southern Methodist University; Leanne Ketterlin Geller, Southern Methodist University; Sarah Powell, The University of Texas at Austin; Erica Lembke, University of Missouri
This study examines the psychometric qualities and initial validity of a measure of teachers' rational number knowledge. Data from 72 teachers revealed good internal consistency. However, item difficulty was not well aligned with teacher abilities. Revisions will focus on validity of test content and improving accuracy in reflecting teacher abilities.

12. Using Bifactor Models to Uncover Construct Validity of Achievement
Junhee Park, University of Iowa; Hyeonjoo Oh, Riverside Insights; Tong Wu
This study examined the construct validity of achievement battery using a bifactor model with Woodcock-Johnson V norming data. Results indicated a good fit, showing that group factors explain additional variance at the test level, highlighting the benefits of the bifactor model in clarifying the achievement construct.

13. Exploring Access to Text-to-Speech Technology's Impact on a Statewide Reading Comprehension Assessment
Annette Vernon, University of Iowa; Catherine Welch; Stephen Dunbar, University of Iowa; Jeongmin Ji, University of Iowa; Dan Song, The University of Iowa; I-Chun Vera Hsiao, University of Iowa
This study examines the validity of a policy decision to allow access to text-to-speech technology on a statewide assessment of reading comprehension for grades 6 and above by exploring the impact on reading comprehension scores through a series of linear and logistic regressions using data from a recent administration.

14. Evaluating Alignment Criteria for Educational Assessments
Eduardo J Crespo Cruz, University of Massachusetts, Amherst; Stephen Sireci, University of Massachusetts Amherst
Alignment studies provide crucial validity evidence based on test content, yet research on alignment criteria is limited. This study explores traditional and alternative alignment criteria using statewide assessment data. Preliminary results suggest that stricter criteria better describe how well tests align with depth of knowledge and content standards.

15. Exploring Effects of Trial By Trial Feedback on Validity of Dyslexia Screening
Wanjing (Anya) Ma; Mia Fuentes-Jimenez; Julian Siebert, Stanford University; Ana Saavedra, Graduate School of Education, Stanford University; Carrie Townley-Flores, Graduate School of Education, Stanford University; Adam Richie-Halford, Graduate School of Education, Stanford University; Benjamin Domingue, Stanford University; Jason Yeatman, Stanford University
This research examines how trial-by-trial feedback impacts the validity of dyslexia screening in a single-word reading assessment, including a randomly ordered Spanish version and an adaptive English version. We analyze students' response behaviors, model fit, and concurrent validity. Findings suggest that feedback enhances student engagement without compromising psychometric validity.

16. Evidence of Fairness Is Essential for Interpreting Engineering Students' Conceptual Understanding
Sandra Liliana Camargo Salamanca, University of Illinois Urbana-Champaign; Kerrie Douglas, Purdue University; Maria Oliveri, Purdue University
This study aims to compare two approaches for detecting gender-based measurement bias in a Concept Inventory (CI) among undergraduate engineering students, particularly with unequal sample sizes. It will also examine how these findings may impact students' average total scores using Measurement Invariance (MI) and Differential Item Functioning (DIF) analyses.

062. Comparing Approaches for Using Large Language Models for Scoring and Specialized Assessments

Individual Paper Session
1:30 PM - 3:00 PM
Hilton Denver City Center Lower Level 2: Denver 5-6

Chair:
Scott Wood, ACT, Inc.; Meng Fan, HUMRRO

Presentations:

Automated Coding with ChatGPT: Success and Challenges
Jiangang Hao, ETS; Wenju Cui, ETS; Patrick Kyllonen, ETS; Emily Kerzabi, ETS; Lei Liu, ETS
We report empirical findings on using ChatGPT to directly code communication data from collaborative tasks. Based on multiple datasets and coding frameworks, we found that ChatGPT is comparable or outperform human raters when the discussions were characterized by colloquial languages but fell short when the discussions dealt with specialized.

Leveraging Large Language Models and Prompt Engineering for Automated Essay Scoring
Yue Huang, Measurement Incorporated; Joshua Wilson, University of Delaware; Corey Palermo, Measurement Incorporated
This study compares a scoring engine using machine learning with large language models using different prompt strategies for automated essay scoring. While the scoring engine had higher agreement with human scores, LLMs provided more consistent results across student groups. Findings highlight the importance of prompt engineering for improving LLM-based scoring.

Leveraging LLMs with Auxiliary Information to Improve Automated Scoring of Science Items
Mingfeng Xue, University of California, Berkeley; Yi-Fang Wu, Cambium Assessment, Inc.
We evaluated the effectiveness of large language models (LLMs) in automated scoring of science assessments, focusing on RoBERTa, DeBERTa, and SciBERT. We assessed performances among these models and the impact of response times as auxiliary information. We found DeBERTa outperforms the others, with some accuracy improvements from incorporating response times.

Use of Natural Language Processing in Automated Scoring of Prosodic Reading Ability
Kuo Wang, Southern Methodist University; Yusuf Kara, University of Miami; Akihito Kamata
Pausing is a typical prosodic feature in assessing prosody, one of the major indicators of oral reading fluency. This study develops a trained classifier-based chunker using natural language processing to split sentences into word chunks. Extensive pausing times related to chunks are then used to predict prosody scores.

Discussants:
Alexander Kwako, Cambium Assessment
Yanlin Jiang, NCCPA

063. Longitudinal/Multilevel Modeling

Individual Paper Session
1:30 PM - 3:00 PM
Hilton Denver City Center Lower Level 2: Colorado IJ

Chair:
Meng Fan, HUMRRO

Presentations:

Enhancing Accuracy in Student Growth Measurement: Adjusting Diagnostic Growth Percentiles
Madeline Schellman, Pearson; Matthew Madison, University of Georgia
This paper introduces three new metrics for measuring relative student growth within a diagnostic modeling framework. These metrics address uncertainty in proficiency diagnoses and introduce penalties for regressing over time, providing more actionable insights compared to other growth measures like student growth percentiles and basic diagnostic growth percentiles.

Predictor Selection in Large-scale Assessments: Penalized Regression Approach for Hierarchical Data
Miryeong Koo, University of Illinois at Urbana-Champaign; Jinming Zhang, University of Illinois at Urbana-Champaign
This study aims to investigate the applicability of a penalized regression approach, glmmLasso, with machine learning technique to large-scale assessment data with multilevel structure for predictor selection. The results show that glmmLasso outperforms Lasso in terms of optimal predictor selection.

Application of cross-classified IRT model to human ratings in longitudinal data
Yun-Kyung Kim, UCLA/CRESST; Eun Hye Ham, Kongju National University
This study applies the cross-classified item response theory model to account for rater influences in a longitudinal context. Using data from a scenario-based assessment administered at three different timepoints and scored by teachers, the proposed approach using the cross-classified IRT model is compared to the widely recognized many-facet rater model.

Filters for IRT Models with Data Collected Across Time
Russell Almond, Florida State University; Andrew Krumm; Blaire Carpenter, Florida State University; Jiawei Li, Florida State University; Elham Tajik, Florida State University
This paper develops time-series filtering methods for unidimensional item response theory (IRT), building analogs of the weighted moving average, the exponential filter and state space models (e.g., the Kalman filter). This method is applied to a data set tracking the performance of surgical residents across time.

The role of reliability when matching prior to a difference-in-difference analysis
Luke Miratrix, Harvard Graduate School of Education; Dae Woong Ham, University of Michigan
In this paper, we explore bias trade-offs when using matching before a Difference-in-Differences (DiD) analysis when the parallel trends assumption is violated. We mathematically show how bias hinges on the reliability of the outcome measure as a function of latent structure in the data. We illustrate with a principal turnover application.

Discussant:
Yanlin Jiang, NCCPA

064. Measuring Complex Skills

Individual Paper Session
1:30 PM - 3:00 PM
Hilton Denver City Center Lower Level 1: Penrose 2

Chair:
Tzu-Chun Kuo, Kaplan North America

Presentations:
AI-Assisted Solutions for Social Emotional Skills Assessment Development
Evelyn Johnson, Riverside Insights; Jennifer Robitaille, Riverside Insights; Emily Parker, Brad Bolender, Finetune by Prometric; Sara Prewett
This presentation explores the use of AI-assisted assessment pipelines for developing Social Emotional and Durable Skills assessments. Through a human-in-the-loop model, AI supports subject matter experts in generating, refining, and reviewing materials. Findings compare traditional and AI-assisted approaches in terms of development time, item quality, and effectiveness.

Assessing What We Value About Critical Thinking: An Innovative Approach
Sundance Zhihong Sun, The University of Melbourne; Bruce Beswick, The University of Melbourne
This study develops the Integrated Critical Thinking Test (ICTT) to assess the multiple dimensions of critical thinking in an integrated way. The validation results indicate that the instrument assesses the intended construct and demonstrates acceptable psychometric properties between samples of Australian (N=233) and Chinese (N=247) higher education students.

Measurement of Two-Dimensional Political Ideology in the United States
Taylor True
The purpose of this study is to construct and validate a measurement model for respondent-level political ideology, using survey data from the 2020 American National Election Survey. Multidimensional item response theory enables a granular view of individual attitudes on various policy issues, providing in-depth insight for further research and practice.

Game-based System Thinking Assessment with Continuous IRT Scoring Model
Xinchu Zhao, Roblox; Sizheng Zhu, Roblox; Matt Emery, Roblox; Yao Xiong, Google; Erica Snow, Roblox; Jack Buckley, Roblox
This study applies continuous Item Response Theory (IRT) models to a novel system-thinking assessment with continuous scores, avoiding the need for categorization. The results from the models showed strong correlations with traditional IRT while preserving all score information and reducing inflated differences between less- and more-represented groups, offering improved accuracy.

Refining Computational Thinking Patterns and Efficiency: A Cross-Country Study from ICILS 2023
Yuan-Ling Liaw, IEA Hamburg; Mojca Rozman, IEA Hamburg
This study examines students' problem-solving strategies in computational thinking tasks, using data from the ICILS 2023, covering 24 countries. Computational thinking involves identifying, analyzing, and solving real-world problems through programming. The research aims to identify students' CT strategies and behavioral patterns, refine efficiency measures, and analyze cross-country differences in performance.

Discussant:
Jessica Andrews-Todd, ETS

065. Teacher Evaluation and School Accountability

Individual Paper Session
1:30 PM - 3:00 PM
Hilton Denver City Center Lower Level 2: Colorado CD

Chair:
Lily An

Presentations:

Are Student and School Evaluations Fair when Instructional Exposure at Testing Varies?
Yeow Meng Thum
Although the accuracy and stability of school growth measures may be improved by judiciously aggregating over time and test subjects, they remain problematic if known variation in instructional exposure at testing is ignored. We illustrate how appropriately projected growth scores support fair evaluations regardless of individual difference in instructional exposure.

Affordances and Tradeoffs of Using Matrix Sampling for State Accountability Testing
Benjamin Shear, University of Colorado Boulder
Matrix sampling is often suggested as a method to reduce the footprint of state testing, but would it be feasible? This paper considers technical and practical issues involved with the implementation of matrix sampling in state testing, emphasizing how the validity and reliability of intended test uses would be affected.

Relationship between Students' Within-year Growth and Teacher's Academic Optimism and Growth Mindset
Parth Soni, Indian Institute of Management Ahmedabad
This study investigated the association of teachers' academic optimism (academic emphasis, self-efficacy, and trust in students and parents), and growth mindset with students' within-year growth. The results of multilevel growth model showed that the teachers' academic emphasis was positively associated with students' growth. The implications of the findings are discussed.

Teacher Evaluation as a Data Use Intervention: A Frame Analysis
John Lane, Michigan Assessment Consortium
Little is known about how student growth data included in recent teacher evaluation systems is used as the source of a data-based decision making for teachers. In this ethnographic study, the authors describe how teachers make sense of and respond to a variety of assessments associated with teacher evaluation.

Discussant:
Megan Welsh

066. Barrier and Challenges for Transitioning Students from AA-AAAS to General Assessments

Organized Discussion
1:30 PM - 3:00 PM
Hilton Denver City Center Lower Level 1: Penrose 1

In 2015, the Every Student Succeeds Act (ESSA) placed a 1% cap on participation in alternate assessments. To meet this requirement, states have been moving students from alternate assessments based on alternate academic achievement standards (AA-AAAS) to general assessments. States have been struggling to meet this requirement. This session will address the challenges and barriers that educators, parents, and students encounter when transitioning students from AA-AAAS to general assessments. To begin, presenters will share research on state AA-AAAS policies, data analyses of characteristics of students who have transitioned from AA-AAAS to general assessments, and educator and parent perspectives on the assessment transition process, as well as key questions for each area. Presenters will provide a foundational overview and background information for each question. After the presentations, attendees will have opportunities to share their expertise and engage in small group discussions that focus on each of the areas of research related to alternate assessments (policy, student characteristics, and educator and parent perspectives). Attendees also will have the opportunity to move between discussion tables to engage with other presenters and attendees on different topics. After the table discussions, the presenters will provide concluding remarks and takeaways from this session.

Chair:
Yi-Chen Wu, National Center on Educational Outcomes, University of Minnesota

Presenters:
Yi-Chen Wu, National Center on Educational Outcomes, University of Minnesota
Mari Quanbeck
Martha Thurlow, University of Minnesota
Deborah Taub, OTL Education Solutions

067. Exploring Effective, Equitable Classroom Assessment to Support Teacher Practice and Student Success (NCME Committee Organized Session)

Organized Discussion
1:30 PM - 3:00 PM
Hilton Denver City Center Lower Level 1: Mattie Silks

Our classrooms and schools need more high-quality and equitable assessments that can support instruction and student learning. With support from the NCME Mission Fund, Van Orman, Riley-Lepo, and McMillan engaged worldwide classroom assessment scholars and K-12 teachers in building and refining a framework for effective, equitable classroom assessment (EECA) to support teacher practice called “A Guiding Framework for Effective, Equitable Classroom Assessment for Teacher Practice and Student Success”. In this organized discussion session, the authors of the Framework join with a teacher, a school administrator, a district administrator, and a classroom assessment scholar to discuss the newest version of the Framework. Presenters will (i) introduce EECA and provide clear

examples, (ii) discuss the current status of using EECA in their classrooms, schools, and systems, and (iii) propose strategies to enhance the use of EECA. Our session will offer important take-away messages to researchers, assessment developers, educators, and administrators, and students who can enact and experience EECA.

Chair:
Duy Pham, New Meridian

Moderator:
Dustin Van Orman, Western Washington University

Presenters:
Dustin Van Orman, Western Washington University
Erin Riley-Lepo, The College of New Jersey
James McMillan, Virginia Commonwealth University
Tara Crewe, Denison Montessori
Natalie Fickes, Denison Montessori
Bryan Drost, Rocky River City School District
Heidi Andrade, University of Albany

068. Testing Standards Revision: Focus on Foundations, Operations, & Psychometrics (Joint NCME/AERA Session)

Organized Discussion
1:30 PM - 3:00 PM
Hilton Denver City Center Lower Level 2: Colorado EF

This is the first of two joint sessions that will update the members of AERA and NCME and provide an opportunity for input on the Standards. Specifically, this session will address the Foundations and Operations sections of the Standards, with a focus on validity, reliability, and precision, as well as test development, scaling, administration, and documentation. The Joint Committee (JC) received its charge on August 21, 2024, and has been meeting with experts, and is working on the Standards revision. In this first session, participants will receive an update on the process and hear from members of the JC regarding their work and any current challenges being addressed by the committee regarding measurement foundations and operations. The majority of the session is intended to hear from participants about their thoughts on the ongoing revision, in reaction to the updates or regarding their own questions, hopes, or concerns for the revision.

Chair:
Michael Rodriguez, University of Minnesota

Moderator:
Kristen Huff, Curriculum Associates

Presenters:
Michael Rodriguez, University of Minnesota;
Kristen Huff, Curriculum Associates
Ye Tong, National Board of Medical Examiners
Andy de los Reyes
Chad Buckendahl, ACS Ventures, LLC
Ellen Forte, edCount, LLC
Laura Hamilton, American Institutes for Research
Qiwei He, Georgetown University
Nathan Kuncel
Cara Laitusis, Center for Assessment
Maria Marquine

Rochelle Michel
Michael Russell, Boston College
Pohai Kukea Shultz, University of Hawaii - Manoa
Stephen Sireci, University of Massachusetts Amherst
Stephen Stark, University of South Florida
Mark Wilson, University of California, Berkeley
Frank Worrell

069. Progress and Pathways: Black Students, Scholars, and Sages in Measurement

Coordinated Paper Session
3:15 PM - 4:15 PM
Hilton Denver City Center Lower Level 2: Colorado EF

The 2025 NCME Annual Meeting theme, "Educational Measurement: In Service of Society," highlights the need for the field to address inequities and better serve diverse populations. This follows a long-sought change in educational measurement, reflected in previous themes such as the 2024 theme, "Reconceptualizing Measurement Theory and Practice to Reduce Inequities," and the 2022 theme, "Turn the Page—The Next Chapter for Educational Measurement." This session builds on these themes by exploring how to create a more equitable and inclusive field, particularly from the perspectives of early-career Black scholars. The first paper investigates the experiences of Black students in educational measurement graduate programs, examining academic backgrounds, enrollment factors, and challenges. The second paper explores the lived experiences of Black psychometricians, identifying barriers and proposing solutions to create more inclusive professional pathways. The third paper offers a culturally responsive retelling of psychometric history by amplifying the contributions of Dr. Lloyd Bond, whose work on fairness and equity has been underrepresented in graduate curricula. This session focuses on actionable steps to address these long-sought changes and advance diversity, equity, and inclusion in educational measurement.

Session Organizer:
Jaylin Nesbitt

Presentations:
Beyond Numbers: Exploring Black Students' Experiences in Educational Measurement Graduate Education
Malitšitso Moteane, Winston-Salem State University; Brianna Hooks-Singletary, University of North Carolina Greenboro; Jaylin Nesbitt; Janine Jackson Haynes

Expanding Pathways: Examining the Lived Experiences of Black Psychometricians in Educational Measurement
Janine Jackson Haynes; Jaylin Nesbitt; Malitšitso Moteane, Winston-Salem State University; Brianna Hooks-Singletary, University of North Carolina Greenboro

A Culturally Responsive Retelling of Psychometric History: Re-reading and amplifying the
Jaylin Nesbitt; Brianna Hooks-Singletary, University of North Carolina Greenboro; Malitšitso Moteane, Winston-Salem State University; Janine Jackson Haynes

Discussant:
Kyndra Middleton, Howard University

070. Research on Remote Proctoring in Support of Valid Test Scores

Coordinated Paper Session
3:15 PM - 4:15 PM
Hilton Denver City Center Lower Level 2: Colorado IJ

As traditional educational models adapt to changing markets and new technologies, ensuring the integrity and fairness of remote assessments is critical for supporting validity arguments of high-stakes tests. To this end, the presentations in this symposium will examine advances and challenges in remotely-proctored assessments, aiming to enhance the testing experience for all stakeholders involved. Central to the discussions is the impact of proctoring on test takers' behavior, performance, and stress levels, as well as the exploration of various methods to mitigate any negative effects. Specifically, the first paper compares test performance across traditional and remote proctoring settings, and proposes a framework to ensure reliable test scores. The second paper introduces innovative measurement models to assess proctoring accuracy and bias in decision making, highlighting their inverse relationship. The third paper evaluates the effectiveness of proctoring through direct observations, while emphasizing the benefits of scrutinizing both in-person and remote exams. The fourth paper examines fairness in remote language testing, and suggests the need for linguistic accommodations to enhance fairness. The final paper proposes a framework for the integration of remote proctoring data as a source of validity evidence, rather than just viewing it as a valuable way to detect malpractice.

Session Organizer:
Alina von Davier, Duolingo

Chair:
Alina von Davier, Duolingo

Presentations:
Measuring Quality and Proctor Decision Making
William Belzak, Duolingo; Yigal Attali; Danielle Mann, Duolingo; J.R. Lockwood, Duolingo

Observation of Exam Administrations to Evaluate and Improve the Effectiveness of Proctoring
David Foster; Marc Weinstein, Caveon

Fairness in Remote Language Testing: The Role of Proctoring Language in Test Taker Experience
Jieun Kim, University of Hawaii at Manoa

Discussant:
Angel Arias, Carleton University

071. How Scoring Decisions Can Introduce Severe Bias into Growth Mixture Model Results

Coordinated Paper Session
3:15 PM - 4:15 PM
Hilton Denver City Center Lower Level 2: Colorado AB

Growth Mixture Models (GMMs) are widely used to identify homogeneous groups based on growth trajectories. However, our recent research reveals that decisions in survey item calibration and scoring can introduce significant bias into GMM results. This coordinated session explores how seemingly minor decisions in scoring can drastically alter the outcomes of GMMs, potentially leading to spurious latent classes and incorrect interpretations of within-class trajectories. Through a series of empirical and simulation studies, we demonstrate the sensitivity of GMMs to various scoring approaches, the effects of unmodeled response styles, and the impact of incorporating covariates during the scoring process. Altogether, these studies have implications for understanding how students develop academically and socio-emotionally, as well as how to identify heterogeneity in the effectiveness of programs supporting that development. The session will provide practical recommendations for researchers to mitigate these biases and ensure more reliable GMM analyses.

Session Organizer:
James Soland, University of Virginia

Presentations:
Why Producing Unbiased Survey Scores for Use in Growth Mixture Models is Almost Impossible
James Soland, University of Virginia

Investigating the Sensitivity of Growth Mixture Models to Conditioning Scoring on Covariates
Stephen Tavares, University of Virginia

Understanding the Effects of Unmodeled Response Styles on Growth Mixture Modeling Results
Veronica Cole, Wake Forest University

Discussant:
Jonathan Templin, University of Iowa

072. Applications of Generative AI to Mathematics Education: Opportunities and Challenges

Coordinated Paper Session
3:15 PM - 4:15 PM
Hilton Denver City Center Lower Level 1: Penrose 2

Large Language Models (LLMs) are optimized for text prediction and generation, and have potential to support many educational applications, from essay coaching to tutoring. However, LLMs sometimes produce inaccurate or imprecise responses, often referred to as "hallucinations." Mathematical domains rely on exact answers and reasoning that can support those answers. Therefore, applying LLMs to mathematics education deserves special scrutiny. The three papers in this session explore if and how LLMs can be applied to mathematics education and mathematics education research. The first paper sets the stage by focusing on the nature of calculation and mathematical reasoning errors LLMs have been observed to make, as well as whether there have been improvements in accuracy over time. The second and third papers both supply the LLM with additional input to guide its efforts; the second paper provides the LLM with a learning progression, and the third paper provides it with a solution graph that captures the steps students take in solving a mathematical problem. The papers raise many questions, such as whether LLMs alone are the appropriate technology to support mathematics education, and if not, whether LLMs can be supported by other technologies.

Session Organizer:
Edith Graf, ETS

Chair:
Edith Graf, ETS

Presentations:
Mathematical Explorations in an LLM
Edith Graf, ETS; Carol Forsyth, ETS; Shona Ruiz Diaz, ETS; Duanli Yan, ETS; Yang Jiang, ETS

Enhancing AI-Math Conversations with Learning Progressions
Diego Zapata-Rivera, ETS; Edith Graf, ETS; Jere Confrey, North Carolina State University, Department of Science, Technology, Engineering, and Mathematics Education; Shona Ruiz Diaz, ETS; Ladislav Stacho, Department of Mathematics, Simon Fraser University

Explorations in using student solution pathways and LLMs for enhanced mathematical learning
Ladislav Stacho, Department of Mathematics, Simon Fraser University; Edith Graf, ETS; Diego Zapata-Rivera, ETS

Discussant:
Mingyu Feng, WestEd

073. Electronic Board Session 2

Electronic Board Session
3:15 PM - 4:15 PM
Hilton Denver City Center Lower Level 2: Colorado GH

Session Organizer:
Ian Campbell, Cambium Assessment

Presentations:

1. **How Adaptive is Your CAT: Measuring Adaptability under Highly Constrained Test Blueprints**
   Ian Campbell, Cambium Assessment
   Large-scale state assessments often require complex test blueprints with many layers of requirements. When administered as CATs, it is often unclear how adaptive these tests are in practice. This study builds on previous literature by exploring adaptability metrics under realistic, highly complex blueprints, demonstrating the benefits of item-level adaptability measures.
2. **Bayesian Classification of Multiple Skills with Multiple Choice Items**
   Sangbeak Ye, Florida Atlantic University; Jimmy de la Torre, The University of Hong Kong
   In computerized adaptive assessments that aim to instruct complex sets of skills, items often require specific skills in the presence of distractors. This research developed an on-the-fly subject skills classification method that utilizes multiple-choice data to provide fine-grained diagnosis and feedback until complete mastery is achieved.
3. **Bayesian Information Theoretic Model-Averaging Stochastic Item Selection in Computer Adaptive Testing**
   Joshua Chang; Edison Choe, Renaissance
   We formulate the optimization problem for CAT in terms of Bayesian information theory, using the statistical distance between the ongoing ability estimate and the full-test ability estimate as a selection metric.
4. **Practical Considerations of Transition to Multistage Test with Limited Item Pool**
   Yanming Jiang, HMH
   This study examines the feasibility of transitioning from a linear test to a multistage test (MST) while preserving a similar estimation precision for an assessment with a limited item pool. It may help practitioners make informed decisions about such a transition.
5. **Dependencies in Item-Adaptive CAT Data and DIF Detection: A Multilevel Framework**
   Danielle (Dandan) Kaptur, Pearson; Justin Kern, University of Illinois at Urbana-Champaign; David Shin Pearson; Jinming Zhang, University of Illinois at Urbana-Champaign
   This study presents a two-level model for detecting differential item functioning (DIF) in computerized adaptive testing (CAT). By accounting for dependencies created by provisional ability estimates, our model outperforms traditional methods in Type-I error control and power. Future research should explore expanding item pools and incorporating item parameters in modeling.
6. **Ensuring Cultural Validity: Challenges and Practices in Developing a Bilingual Attitudinal Survey**
   Diley Hernández, Georgia Institute of Technology; Jayma Koval, Georgia Institute of Technology
   This study reflects on the challenges, opportunities, and lessons learned in developing and implementing a bilingual (English-Spanish) and culturally valid attitudinal survey for Latine youth in the U.S. and Puerto Rico. It shows practical examples of how to address sociocultural, linguistic, epistemological, contextual, and developmental validity.
7. **Validity and Service to Society**
   Josh Lederman, Brandeis University
   This paper advances a concept of validity that centralizes "the 'why' of our work in educational measurement" as the betterment of society. The author provides an argument-based model of validity and validation that forefronts societal good as the explicit purpose of assessment, and validity as the actualization of such good.

8. IRT-based Method for Detecting Low Discrimination Items
Yuefeng Tang, College Board; Weiwei Cui
We propose two IRT-based methods for detecting low discrimination items in the framework of IRT accounting the coverage of samples for IRT models. Results show that the two proposed methods are robust to detect items potentially having low discrimination and are powerful to identify difficult items with low discrimination.

9. Impact of Item Parameter Drift and DIF on IRT Number-Correct Scoring
Anthony Fina, University of Iowa; Jing Ma, University of Iowa
The impact of IPD and DIF on IRT number-correct scoring is examined and contrasted with EAP scoring. Preliminary evidence suggests that number-correct scoring is more robust to IPD, DIF, and violations of IRT assumptions. The results from this study also provide new insights into how IPD impacts MST performance.

10. Linking Errors Introduced by Rapid Guessing in Multigroup Concurrent IRT Scaling
Jiayi Deng, Human Resources Research Organization
This study investigated the effect of rapid guessing (RG) on multigroup concurrent IRT calibration, a linking method used to ensure comparable scores across different linguistic test forms. While the calibration process generally showed robustness in anchor identification and ability estimation accuracy, RG showed potential to influence the ability estimation precision.

11. Revisiting Anchor Item Evaluation Criteria
Ying Lu, The College Board; Xiuyuan Zhang, The College Board
Various methods have been used for evaluating anchor items and identifying outliers in common-item nonequivalent group equating. This study investigates and compares the stringency of the Mantel-Haenszel procedure and the IRT-based item characteristic curve (ICC) procedure that evaluates the differences in ICCs through RMSE and MAD.

12. Leveraging Swarm Intelligence to Detect Rapid Guessing in Low-Stakes Assessments
Yu Bao, James Madison University; Wei Zhang, Augusta University; Jin Liu, The University of Texas at Arlington
This study investigates the application of swarm intelligence, specifically Standard Particle Swarm Optimization (SPSO), to identify rapid guessing behaviors in low-stakes assessments. Through empirical and simulation studies, we explore how the data-driven response thresholds perform compared to the canonical methods, offering insights into improving test score validity.

13. Are Group Decisions Fair? The Analysis of Residuals between Individual and Group Ratings
Stella Kim, University of North Carolina Charlotte; Lisa Keller, University of Massachusetts, Amherst; Jonathan Supovitz,University of Pennsylvania
This study aims to examine the differences between the two evaluation methods (individual ratings and group ratings) and identify potential factors affecting group-level decisions.

14. Incorporating Response Time in Early Mathematics Assessments: A Hierarchical Modeling Approach
Xiaochen Xu, University of California, Davis; Tony Albano, University of California - Davis; Robin Hojnoski, Lehigh University; David Purpura, Purdue University; Kristen Missall, University of Washington
This study applies hierarchical models to early mathematics assessment data, incorporating response time using the Rasch and exponential models. By exploring the relationship between response time and accuracy, we aim to improve item calibration and proficiency estimation. Results will contribute to psychometric advancements in early education assessment.

15. Effect Size Impacts on DIF Detection: Item-Focused vs. Rasch Trees
Nana Amma Asamoah, University of Arkansas; Ronna Turner, University of Arkansas
This study examines the impact of incorporating effect size measures in applications of item-focused and Rasch trees for DIF detection. Initial findings reveal the inclusion of effect sizes improves DIF detection accuracy for item sets, increasing correct DIF solutions and reducing the occurrence of solutions including incorrect DIF detection.

16. **The Exploratory Structural Equation Modeling Approach for the Coparenting Relationship Scale Adaption**
Fajer Shamsaldeen, Kuwait University
Most cross-cultural scale adaption studies rely on exploratory or confirmatory factor analysis (EFA/CFA). This study aims to use the Exploratory Structural Equation Modeling (ESEM) to assess the factorial validity of the Coparenting Relationship Scale in Arabic-speaking populations. The ESEM overcomes the limitations of EFA and CFA by incorporating their advantages.

074. Using AI to Address Challenges in Scoring and Automated Feedback

Individual Paper Session
3:15 PM - 4:15 PM
Hilton Denver City Center Lower Level 2: Denver 4

Chair:
Teresa Ober, ETS

Presentations:

A Framework for Evaluating Automated Conventions Feedback
Mackenzie Young; Amy Burkhardt, Cambium Assessment; Quinell Wilson, Cambium Assessment; Susan Lottridge, Cambium Assessment
While AI tools can support learners by providing automated feedback to improve writing conventions, it is equally important to critically evaluate the quality of that feedback. This paper introduces a framework for such evaluation, shares a benchmark dataset, and applies the framework to a recently developed AI feedback tool.

Automated Scoring of Long Essays
Kuo Wang, Southern Methodist University; Haowei Hua, The Culver Academies; Pengfei Yan, University of Maryland, College Park; Hong Jiao, University of Maryland; Dan Song, The University of Iowa
Large Language Models (LLMs) are currently the state-of-the-art for automated scoring. However, some LLMs may fall short for automated scoring of long essays with over 10,000 words. Thus, this research evaluates the performance of LLMs, their ensemble techniques, and generative language models GLMs for scoring long essays.

Automated Essay Scoring with Large Language Models: An Evaluation of Different Modeling Approaches
Rudolf Debelak, University of Zurich; Laura Stahlhut, Institut für Bildungsevaluation Zürich AG; Alexandre Clin Deffarges, Institut für Bildungsevaluation Zürich AG; ETH Zurich
This study compares generative and classical machine-learning based models for automated essay scoring in a sample of English essays. While classical machine learning-based models show a higher agreement with human ratings, generative models offer the potential of slightly less accurate assessments with less training data.

075.

Discussant:
Ummugul Bezirhan, Boston College

Dimensionality/Factor Analysis

Individual Paper Session
3:15 PM - 4:15 PM
Hilton Denver City Center Lower Level 2: Denver 1-2

Chair:
Hyunjung Lee, Fordham University

Presentations:

Rank Factor Score as a Robust Scoring Method in Factor Analysis
Jihong Zhang, University of Arkansas at Fayetteville; Xinya Liang, University of Arkansas; Ji Li,

University of Arkansas
This study develops a robust scoring method for estimating latent factor scores using rank-based factor loadings. By ranking loadings, the method focuses on the relative strength of item-factor relationships, which improves factor score estimation in situations where the models are misspecified or when extreme values distort the estimates.

A Framework for Extracting Latent Dimensions from Response Time Data
Guoliang Xu, Teachers College, Columbia University; James Corter, Teachers College, Columbia University
This paper proposes a method based on individual differences multidimensional scaling (INDSCAL), also known as weighted MDS, to extract meaningful latent dimensions from response timing data in computer-delivered tests, enhancing predictive validity and assessment precision compared to using items scores only.

Careless Responding May Yield Spurious Factors in Dimensionality Assessment: A Simulation Study
Mohammed Abulela, University of Minnesota; Kyle Nickodem, University of Minnesota - Twin Cities; Michael Rodriguez, University of Minnesota
Careless responding (CR) has been found to negatively impact validity evidence. Yet, its potential to yield spurious factors in dimensionality assessment has not been comprehensively investigated. Accordingly, we conducted a simulation study where CR type, prevalence, and severity were manipulated. We found CR yielded a second factor for unidimensional data.

Discussant:
Jonathan Weeks, American Institutes for Research

076. IRT Item Calibration

Individual Paper Session
3:15 PM - 4:15 PM
Hilton Denver City Center Lower Level 2: Denver 5-6

Chair:
Andrea Gotzmann, Medical Council of Canada

Presentations:
Rethinking IRT Calibration with New Algorithmic Approaches
Ruitao Liu, 2U
This study introduces the Simulated Gradient Ascent Method to address limitations in traditional Item Response Theory calibration. SGAM adapts to non-normal ability distributions, offers individualized calibration, and improves computational efficiency. A simulation study compares SGAM with traditional methods, evaluating its effectiveness in IRT model calibration for diverse, large-scale assessments.

Impact of Different IRT Calibration and Scaling Methods on Vertical Scale
Hyeonjoo Oh, Riverside Insights; Hanwook Yoo, Ascend Learning; Tim Moses, Buros Center for Testing
This study investigates the impact of different item response theory (IRT) calibration methods and scaling constant calculations on vertical scales using simulated data. Preliminary results indicate that, overall, the 2PL summed scoring condition resulted in smaller RMSDs compared to other simulation conditions.

Evaluating Calibration Methods for Embedded Field Test Items
Shichao Wang; Chunxin Wang, ACT, Inc.; YoungWoo Cho, ACT
This study explores the efficacy of three calibration methods for embedded field test items using real data and simulated data under various conditions. It contributes to the ongoing redesign of the item calibration processes, ensuring that educational assessments maintain the highest standards of measurement accuracy and reliability across administrations.

Discussant:
Ed Merkle, University of Missouri

077. Modeling Process Data

Individual Paper Session
3:15 PM - 4:15 PM
Hilton Denver City Center Lower Level 2: Colorado CD

Chair:
Suhwa Han, Cambium Asessment

Presentations:

Automated Monitoring of Test-taking Behaviors in Online Assessments
Suhwa Han, Cambium Asessment; Hyeon-Ah Kang, University of Texas Austin
This study presents fully-automated methods for real-time monitoring of examinees' test-taking behaviors. Without prior assumptions on examinees or items, the methods identify anomalies by applying machine learning to features leveraging measurement models on process data. The methods have been applied to various simulated data and a state-wide summative assessment.

Latent Markov Models for Multivariate Time-Series Data From Computerized Assessments
Hyeon-Ah Kang, University of Texas Austin
The study proposes a refinement framework for the latent Markov model (LMM) that performs interaction analytics on computer-logged assessment data. We extend the current development in LMM to support latent sequence analysis of multimodal time-series and enable inference on the event-specific measurement parameters.

The Impact of Students' Disengagement in Untimed Tests on Predicting Future Success
Tzu-Chun Kuo, Kaplan North America; Nina Deng, Kaplan North America; Snigdha Rai, Kaplan North America
An untimed test can cause a positively skewed response time as students spend as much time as needed. This study examines the impact of skewed response time on non-effortful behavior identification and predictive model fit. The result showed that students with more disengaged responses had a lower probability of success.

Discussant:
Constanza Mardones-Segovia, University of Georgia

078. Score Reporting

Individual Paper Session
3:15 PM - 4:15 PM
Hilton Denver City Center Lower Level 2: Denver 3

Chair:
Jessalyn Smith, DRC

Presentations:

Do Not Associate Red With Our Children: Somali Perspectives on Score Reporting
Samuel Ihlenfeldt, University of Minnesota
When score reports are not translated, many parents cannot equitably access the resources to understand their childrens' test scores. A series of focus groups were conducted in Somali to explore parents' perspectives on the design and content of adapted English and translated Somali score reports. Implications and recommendations are discussed.

How Many Observations are Needed to Identify Problematic Human Raters?
Yangmeng Xu, University of Alabama; Edward Wolfe, Pearson
Scoring efficiency and accuracy are critical concerns when evaluating raters in constructed-response assessments. We examined the efficiency and accuracy of raters' scoring practices and the

use of statistical indicators for identifying problematic raters. We identified a minimum number of student responses to reach acceptable accuracy.

Added Value of Subscores for Tests with Polytomous Items
Kylie Gorney, Michigan State University; Sandip Sinharay, ETS
Sinharay (2010) concluded that it is not common for subscores to have added value. However, Sinharay almost exclusively dealt with data from tests with only dichotomous items. In this paper, we show that it is more common for subscores to have added value in tests with polytomous items.

Discussant:
Marianne Perie, Measurement in Practice

079. Generative-AI automatic item generation for large-scale language assessments

Coordinated Paper Session
4:30 PM - 6:00 PM
Hilton Denver City Center Lower Level 2: Colorado AB

Internet-based test administration offers numerous advantages over traditional paper-based methods, such as adaptive item administration, automated scoring, and immediate feedback. These benefits have increased both summative and formative testing, leading to a higher demand for item development. Automatic item generation (AIG) addresses this need by creating test items using predefined models or templates. Despite its utility, AIG was traditionally limited to simpler tasks and required heavy reliance on expert input, increasing costs and resource demands. Complex tasks, such as those assessing reading comprehension, are challenging for traditional AIG because they require unique items tailored to specific passages. However, recent advancements in generative AI could potentially revolutionize AIG, and test development in general, for more sophisticated assessments.

In this session we present a new paradigm for test development of complex tasks using generative AI. This paradigm allows a fully automated process for the creation of complex assessments, such as reading and listening comprehension tasks, including source passages and a wide range of comprehension questions. Moreover, the paradigm has been applied to the creation of innovative writing and speaking production tasks that are interactive in real-time, allowing us to improve the measurement of complex constructs in large scale assessment.

Session Organizer:
Yigal Attali

Chair:
Yigal Attali

Presentations:
Principles For Generative-Ai Assessment Task Generation
Andrew Runge, Duolingo

Human-Centered Quality Assurance For Automatically-Generated Assessments
Jacqueline Church, Duolingo

Iterative Evidence-Based Task Design For Generative Ai Assessments
Yena Park, Duolingo

Theme-Based Follow-Up Prompting For Automatically-Managed Interactive Essay Writing Assessment
Sarah Goodwin, Duolingo

An Automatically-Managed Adaptive Interactive Speaking Assessment
Yigal Attali

Discussant:
Christopher Ormerod

080. Examinee Preknowledge: Understanding the Impact, Mitigating the Implications, and Detecting the Imprint

Coordinated Paper Session
4:30 PM - 6:00 PM
Hilton Denver City Center Lower Level 1: Penrose 2

This coordinated session brings together research addressing critical challenges in maintaining test security, exploring the effects of compromised items (CI) and examinees with preknowledge (EWP), and moving beyond detection to investigate the broader impact on examinee scores and item parameter recovery. We cover two broad areas: (i) the effects of preknowledge on parameter estimates and downstream decisions and (ii) intuitive detection methods for EWP-CI. First, we examine how bias introduced into proficiency estimates by preknowledge impacts classification accuracy for (non) EWP. Second, we investigate how multidimensional IRT can mitigate this parameter estimate bias. For detection, we propose an iterative method to identify EWP and CI simultaneously using anomaly detection and chi-square tests of independence, providing a model-free alternative to more complex approaches. Lastly, methods for detecting preknowledge-driven collusion are reviewed, with a focus on response similarity and clustering, offering strategies tailored for small-scale assessments that may be underpowered for traditional techniques. This session emphasizes practical approaches to both understanding and addressing the risks of preknowledge, aiming to provide researchers and practitioners with insights into minimizing its impact and safeguarding both the integrity of testing programs and the validity of inferences made on the basis of their scores.

Session Organizer:
Nate Smith, University of Massachusetts, Amherst

Presentations:
Impact of Compromised Items and Examinees With Preknowledge on Proficiency Score Estimates
Kevin O'Rourke, University of Massachusetts Amherst; Rob Keller, HumRRO

Course Correction: Mitigating Item Parameter Bias from Preknowledge with Multidimensional IRT
Andrea Hebert, Cognia; Nate Smith, University of Massachusetts, Amherst

Do the Two-Step: Iterative Detection of Compromised Items using Tests of Independence
Nate Smith, University of Massachusetts, Amherst; Lisa Keller, University of Massachusetts, Amherst

Collusion Detection in Small-Sample Assessments: Review of Pairwise and Group Similarity Analyses
Richard Feinberg, NBME

Discussant:
Sandip Sinharay, ETS

081. Item Bank Maintenance in Three-Dimensional Science Assessment Programs

Coordinated Paper Session
4:30 PM - 6:00 PM
Hilton Denver City Center Lower Level 2: Colorado CD

This session proposes a comprehensive item bank maintenance plan. We propose that the items in a bank are periodically reviewed by content experts, and re-field-tested to obtain updated item statistics and item parameters. Our approach can be contrasted with a reactive model where items are re-field-tested only after being flagged for, for example, item drift. While we propose that item parameters are always updated so that they are based on the most recent data, the re-field-test design also allows to investigate potential item parameter drift (IPD) and the factors that may cause it like curriculum shifts or test environment characteristics. We piloted the item bank maintenance design for a large item bank for three-dimensional science. Science content specialists discuss how items are systematically evaluated for content. We selected a portion of our operational item bank

for re-filed-testing and utilized a spiraled delivery design. Our study compares the initial calibration of field-tested (FT) items with re-field-tested (re-FT) items to assess drift. Additionally, the study investigates factors contributing to IPD, including item content, format, and student characteristics such as proficiency and response times. Findings provide actionable insights into maintaining item bank integrity and ensuring the validity of test scores across administrations.

Session Organizer:
Allison Boykin, Cambium Assessment

Presentations:
Item Selection and Design
Mengyao Cui, Cambium Assessment; Wanchen Chang, Cambium Assessment

Pilot Design
Frank Rijmen, Cambium Assessment

Content Approach to Item Drift
Heather MacRae

Calibration and Detecting Drift
Allison Boykin, Cambium Assessment

Exploring Factors Related to Item Parameter Drift
Mina Lee, Cambium Assessment; Wanchen Chang, Cambium Assessment

082. Using AI and Machine Learning to Improve Data on Students' Educational Progress in the US and Worldwide

Coordinated Paper Session
4:30 PM - 6:00 PM
Hilton Denver City Center Lower Level 2: Denver 3

National and international large-scale survey assessments (LSAs) of students' educational achievement, e.g., NAEP in the US or PISA or TIMSS internationally, provide critical information on education systems. For example, 2022 survey results NAEP and PISA contained vital evidence of the impact of COVID on students' knowledge and their initial recovery worldwide. For 30+ years, these surveys have relied on a common statistical approach to evaluate the ability of student populations and human raters to score constructed responses to open-ended items. Machine learning (ML), artificial intelligence (AI), and increased computing power offer opportunities for rethinking the statistical methods used in LSAs. The session presents four papers that offer alternatives to the tried-and-true methods that have been the foundation of LSAs. Three papers rely on technology - ML, AI, and increased computing power drive their alternative methods. Two offer alternatives to the current 'plausible value' approach to estimating ability, which may yield more accurate or informative results and greater flexibility for secondary analysis. The third uses AI instead of humans to score constructed responses in multiple languages. The final paper proposes an alternative to the large-scale, random-assignment bridges used to ensure comparability with historical scores whenever changes are introduced to the LSA procedures.

Session Organizer:
Daniel McCaffrey, ETS

Chair:
Daniel McCaffrey, ETS

Direct Estimation: A Flexible Approach for Large-Scale Assessment Data with Plausible Values
Leslie Rutkowski, Indiana University

Improving Open-ended Scoring in TIMSS and PIRLS using LLMs
Matthias von Davier, Boston College

Cost-Efficient Approaches to Measurement Changes in Latent Variable Modeling
Paul Jewsbury, ETS; Daniel McCaffrey, ETS

Discussant:
Daniel McCaffrey, ETS

083. Electronic Board Session 3

Electronic Board Session
4:30 PM - 6:00 PM
Hilton Denver City Center Lower Level 2: Colorado GH

Session Organizer:
Seong Eun Hong, Center for Applied Linguistics

Presentations:

1. **Exploring Speed-Accuracy Trade-Off in Multistage Adaptive Language Assessment**
   Seong Eun Hong, Center for Applied Linguistics; Yage Guo, Center for Applied Linguistics; Shu Jing Yen, Center for Applied Linguistics
   This study investigates the relationship between within-person variation in response time and accuracy across different age groups and proficiency levels using the multistage adaptive language assessment. The preliminary results show that the speed-accuracy curves vary by age groups and language domains.
2. **Investigating Rapid Responding and Response Time in a Multistage Adaptive Language Assessment**
   Yage Guo, Center for Applied Linguistics; Seong Eun, Hong Center for Applied Linguistics
   Rapid responding is a common challenge in adaptive testing environments, especially in language assessments. Rapid responding has deleterious effects on both measurement properties and score-based inferences. This study explores the relationship between response time and rapid response across different age groups and proficiency levels in a multistage adaptive language assessment.
3. **Adaptively regulated interim theta bounds for optimizing CAT item selection**
   Sung-Hyuck Lee, GMAC; Kyung (Chris) Han, GMAC
   With the increasing preference for shorter tests, strategic item selection becomes crucial in a CAT. However, variability in interim theta poses a significant issue in the early stage of a CAT. This study highlights the need to regulate interim theta estimates for more optimal item selection in shorter tests.
4. **Direct Scaling Methods to Calibrate Pretest Items in Multistage Testing**
   TsungHan Ho, ETS
   The Bayesian logistic regression (BLR) is a direct scaling method developed to address the response sparseness of linking items leading to concern about the quality of the link to the item bank in multistage testing. Two enhanced BLR variations are proposed to improve the performance evaluated using the simulation data.
5. **IRT True- and Observed-Score Equating for Mixed-Format Tests with Weights**
   Ying Lu, The College Board; Xiuyuan Zhang, The College Board; Min Wang, College Board; Lei Wan, College Board
   Research on IRT equating is limited for mixed-format tests with item/section weights. This study aims to compare IRT true and observed score equating with equipercentile equating and offer considerations for selecting appropriate equating method for mixed-format tests with weights.

6. Psychometric Modeling to Identify Examinees' Strategy Differences During Testing
Clifford Hauenstein, Johns Hopkins School of Medicine; Susan Embretson, Georgia Institute of Technology; Eunbee Kim, Georgia Institute of Technology
We propose a set of time series measurement models that identify construct-distant response processes by simultaneously considering both unexpectedly long and unexpectedly short response times.We also demonstrate the potential for such an approach to serve both validity check and correction functions with an empirical example.

7. Exploring Item-Level Confidence Ratings using Explanatory Item Response Models: Illustration and Simulation
Stefanie Wind, The University of Alabama; Randall Schumacker, The University of Alabama
We consider how different specifications of explanatory item response models can be used to model item-level confidence ratings as an item-covariate, person-covariate, and in a combined model. We demonstrate the three approaches using a real data analysis and explore broader applicability using a simulation study.

8. An alternative response times model for survey questionnaires
Hui-Fang Chen, Department of Social and Behavioural Sciences, City University of Hong Kong
The present study proposed not to remove outliers of response times (RTs), as they may indicate varied responding behaviors in survey questionnaires, and to conduct mixture IRT models to explore latent membership. Then latent regression approach can be conducted to examine the relationship between latent class membership and responding patterns.

9. Providing Evidence of Content Validity Using Latent Semantic Analysis
Richard Brown, West Coast Analytics; Sarah Murphy Council For Professional Recognition
This study explores the application of latent semantic analysis to identify the degree of alignment between the standards of the Child Development Associate (CDA) credential, which were derived from a job task analysis of early childhood educators and professionals, and a collection of Early Childhood Educator (ECE) standards.

10. An Empirical Comparison of Linking Methods in Multistage Adaptive Test
Hacer Karamese, WIDA at the University of Wisconsin-Madison; Kyoungwon Bishop, Sooyong Lee; WIDA at the University of Wisconsin-Madison; Syed Abdul, Hadi
This study examines procedures for linking old and new scales in a large-scale multistage adaptive testing program after establishing a new vertical scale, and it discusses the practical implications of these methods. The study conditions included different data collection designs, calibration methods, smoothing, linking group, and group invariance.

11. Developing a Spanish Sentence Reading Efficiency Measure Fair for Multilingual Learners: ROAR-Frase
Mia Fuentes-Jimenez; Wanjing; (Anya) Ma; Julian Siebert, Stanford University; Ana Saavedra, Graduate School of Education, Stanford University; Carrie Townley-Flores, Graduate School of Education, Stanford University; Adam Richie-Halford, Graduate School of Education, Stanford University; Jason Yeatman, Stanford University
This study focuses on development and validation of an online sentence reading efficiency assessment in Spanish (Frase). Frase demonstrates high reliability in monolingual and multilingual populations providing fair, actionable scores. This assessment can be used to gain an understanding of multilingual learners' foundational reading skills on a normative monolingual scale.

12. A Comparison of Bayesian and Maximum Likelihood Estimation in Bifactor and Bifactor Predictive Models
Xinchang Zhou; Bo Zhang, University of Illinois at Chicago; Yan Xia, University of Illinois at Urbana-Champaign
This study employs Monte Carlo simulations to evaluate parameter estimates and fit indices of Bayesian estimation versus maximum likelihood estimator (MLE) in the context of bifactor and

bifactor predictive models. Results show that Bayesian estimation outperformed MLR in small sample sizes, particularly in reducing bias and improving convergence rates.

13. **Analyzing the Dimensional Structure of OSCEs with a Multidimensional IRT Model**
Nai-En Tang, National Board of Chiropractic Examiners; Igor Himelfarb
This study assesses four competing IRT models—two-dimensional, bifactor, testlet, and two-tier models—to determine their suitability for the chiropractic OSCE license exam. Our analysis reveals that the bifactor model offers the best fit for the exam data, highlighting the existence of a complex structural dimension within the OSCE.

14. **Measurement Comparability by using Response Time in Subgroup**
Hsin-Ro Wei, Riverside Insights; Hyeonjoo Oh, Riverside Insights
This study explores test-taking behavior using differential response time (DRT) analysis. By comparing the response times of different subgroups on a cognitive assessment, the research highlights the impacts of DRT on measurement comparability. DRT analysis aids in identifying potential biases, supporting fairness evaluations.

15. **Tracking Learning Over Time Using Cognitive Diagnosis Models**
Wenjing Guo, Pearson; John Binzak, Pearson Clinical Assessments; Tianshu Pan
We used cognitive diagnosis models (CDM) on longitudinal data to examine changes in students' learning over time. Our findings demonstrate that CDMs are effective in tracking learning progress and identifying shifts in specific attribute. This information allows educators to customize instruction and address achievement gaps based on ethnicity-related learning differences.

084. Bayesian Inference for Latent Variable Modeling

Individual Paper Session
4:30 PM - 6:00 PM
Hilton Denver City Center Lower Level 2: Denver 5-6

Chair:
Tao Wang, University of Victoria

Presentations:

Posterior Predictive Model Checking in Bayesian Structural Equation Modeling: Integrating Kullback-Leibler Divergence
Ejike Edeh
This study proposes a reformulated Posterior Predictive P-value (PPp) by replacing the Chi-Square discrepancy function, as currently implemented in commonly used Bayesian Structural Equation Modeling analysis software, with Kullback-Leibler divergence. It hypothesizes that this modification will better capture model fit, offering improved accuracy over current Chi-Square-based PPp index.

Investigating the Performance of Bayesian Model Fit Indices in Ordinal Confirmatory Factor Analysis
Tiejun Zhang; Xiaobo Wei, University of South Carolina; Fang Wang, University of South Carolina
This study aims to examine the performance of Bayesian model fit indices (e.g., PPP, BCFI, BTLI, BRMSEA) in comparison to frequentist indices (e.g., MLR and WLSMV-based ?□, CFI, TLI, RMSEA) across different sample sizes, degrees of item skewness, and levels of model misspecification in ordinal Confirmatory Factor Analysis (CFA).

Extending the Confirmatory Item Set Response Model to Polytomous Items
Victoria Quirk, University of Illinois at Urbana-Champaign; Justin Kern, University of Illinois at Urbana-Champaign
We expand on the work done by Kern & Culpepper (2020), and Kern (2024) to extend the confirmatory item set response model (ISRM) to allow for polytomous items. By allowing for polytomous item structure, we increase the potential applications of this model to psychological and other non-cognitive assessments.

Examining Sources of Illusory Halo Effects with Mixture Rasch Facets Models
Kuan-Yu Jin, Hong Kong Examinations and Assessment Authority; Thomas Eckes, University of Bochum
Recent studies on illusory halo effects focus on the inability of halo raters to distinguish between different rating criteria, ignoring the possibility that certain ratee characteristics may induce illusory halo. In this study we expanded the Rasch facets model to consider these two mechanisms and estimate parameters using the Bayesian approach.

Discussant:
Luke Miratrix, Harvard Graduate School of Education

085. DCM Q-Matrix Methods

Individual Paper Session
4:30 PM - 6:00 PM
Hilton Denver City Center Lower Level 2: Colorado IJ

Chair:
Yi-Fang Wu, Cambium Assessment, Inc.

Presentations:

Data-Driven Q-Matrix Estimation in Cognitive Diagnosis Models Using Machine Learning
Jujia Li, University of Alabama; Kaiwen Man, University of Alabama; Joni Lakin, University of Alabama; Wei Huang, Universtiy of Alabama, College of Education
This study explores machine learning methods for conducting the Q-matrix in the GDINA model, comparing ML approaches to traditional methods (e.g., PCA and GDI). Through simulations across conditions, we assess model fit, parameter accuracy, recovery rates, and the impact of different sample sizes and attribute/item ratios on performance.

Methodological Advances in Q-Matrix Refinement and Model Selection for Cognitive Diagnosis
Youn Seon Lim, University of Cincinnati
This study presents a methodological framework for refining Q-matrices in cognitive diagnosis models using data-driven techniques. We detail the application of the General Diagnostic Index and mesa plots, demonstrating how these methods enhance model fit and diagnostic accuracy in educational assessments, with a case study in biochemistry.

A Novel Parametric Bootstrap Approach for Detecting Testlet Effects in Cognitive Diagnosis Models
Youn Seon Lim, University of Cincinnati
This study introduces an innovative framework for detecting testlet effects in cognitive diagnosis models using a parametric bootstrap Mantel-Haenszel statistic. By evaluating dependencies at the testlet level, this approach enhances diagnostic accuracy, offering a comprehensive solution to a critical challenge in educational assessments that employ testlet-based formats.

The Impact of Q-matrix on Item Parameter and Attribute Profile Estimations in Transitional Diagnostic Classification Models
Olasunkanmi Kehinde, Elizabeth City State University; Shenghai Dai, Washington State University; Brian French, Washington State University
Longitudinal Diagnostic Cognitive Models (DCMs) have been gaining attention in diagnosing learning progressions. The current simulation study investigates the impact of Q-matrices with hierarchical cognitive attributes on the performance of transitional DCM (TDCM). Results showed that different Q-matrix design significantly impacted item parameter and attribute profile estimation accuracy of TDCM.

Discussant:
Andre Rupp, Center for Assessment

086. Algorithms for Test Development

Individual Paper Session
4:30 PM - 6:00 PM
Hilton Denver City Center Lower Level 2: Denver 1-2

Chair:
Shalini Kapoor

Presentations:

Building an Automatic Rubric Validation System: Assertion Recommender
Jiajun Xu, Cambium Assessment; Suhwa Han, Cambium Asessment; Wei Schneider, Cambium Assessment; Zebing Wu, Cambium Assessment
The paper introduces the "Assertion Recommender," a tool that automates the creation and optimization of scoring rubrics for NGSS assessments. By evaluating multiple criteria simultaneously, the tool assists content experts in validating rubrics through real-time comparisons. This process leads to more informed decision-making, ultimately enhancing the quality of item development.

Rubric Validation of Cluster-Type Items using A Decision Tree Algorithm
Wei Schneider, Cambium Assessment; Jiajun Xu, Cambium Assessment; Suhwa Han, Cambium Asessment; Zebing Wu, Cambium Assessment
This study used decision tree algorithms to statistically identify the knowledge structure of item clusters, aiming to inform the quality of scoring rubrics. The results suggested that decision tree models could effectively differentiate high-quality versus low-quality scoring rubrics, as evidenced by differential prediction accuracy, feature selection, and information gain.

A Genetic Algorithm for Automated Assembly of Linear Multidimensional Forced-Choice Questionnaires
Jianbin Fu, ETS; Patrick Kyllonen,ETS
The paper improves and extends a genetic algorithm, the node histogram-based sampling algorithm (NHBSA), to assemble linear multidimensional forced-choice questionnaires with any block sizes. Based on simulated and real data, NHBSA performs better in maximizing the average reliability of latent scores than the mixed integer programming and randomly assembled forms.

Basic Feature of Distractor Cascade Analysis: Item Characteristics and Entropy
Takahiro Terao, The National Center for University Entrance Examinations
This study aimed to examine the relationship between visual features of distractor cascade analysis and difficulty, discrimination, guessing and testlet effects. Results of the comparison of entropy for selection patterns across multiple items showed that the main source of cascade simplicity came from higher discrimination and lower random guessing.

Discussant:
Matthew Gushta, MetaMetrics, Inc.

087. Bridging Graduate Students and Practice: Internships in Educational Measurement (SIGIMIE-Organized Session)

Organized Discussion
4:30 PM - 6:00 PM
Hilton Denver City Center Lower Level 2: Colorado EF

This Educators of Measurement sponsored session focuses on the importance of internships for master's and doctoral students, examining how these experiences foster professional development and contribute to the field. The session is divided into two parts: a presentation of a research study by Autumn Wild, Joseph Kush, and Brian Leventhal on educational measurement internship offerings,

followed by a panel discussion, featuring varying perspectives of internship stakeholders and a graduate student who has recently participated in an internship program. Findings of the study illustrate (a) the current landscape of internship offerings, (b) a successful intern applicant profile, and (c) internship success from multiple perspectives. The panel discussion will explore the value of internships, featuring insights from Luz Bay, PhD (College Board), Hailey Kuang, PhD (Florida State University), Scott Marion, PhD (National Center for the Improvement of Educational Assessment), and Yichong Cao (Graduate Student at the University of Iowa). Throughout the session, we highlight an alignment between internship opportunities and the NCME Foundational Competencies in Educational Measurement (2023). This session welcomes students, internship organizers, and professionals, offering valuable insights and practical advice for enhancing the internship experience and its benefits for all parties involved.

Chair:
Brian Leventhal, James Madison University

Moderator:
Autumn Wild

Presenters:
Autumn Wild; Luz Bay, College Board
Yichong Cao, The University of Iowa
Huan Kuang, Florida State University
Scott Marion, Center for Assessment

088. One Hundred Years of Criticisms Against Standardized Testing

Organized Discussion
4:30 PM - 6:00 PM
Hilton Denver City Center Lower Level 1: Penrose 1

In this session, we will engage in a comprehensive debate about the history of criticisms against standardized testing and explore potential paths for creating a sustainable future for the field of educational measurement. We will examine a collection of historical documents that feature both criticisms and defenses of standardized testing, providing a broad overview of how this technique has been perceived and challenged over time. These documents, curated by Dr. John Fremer, offer valuable insights into key debates surrounding the fairness, effectiveness, and societal impact of standardized assessments. Through the analysis of these records, we will identify recurring themes and pose critical questions to the panelists regarding their views on the legitimacy of these criticisms. Panelists will be asked to reflect on how these historical critiques remain relevant today and what changes might be necessary to ensure that standardized testing evolves in ways that promote equity and address past shortcomings. The session will also consider the broader implications of these criticisms for the future of educational measurement. Ultimately, we aim to chart a constructive path forward for the field.

Chair:
Sergio Araneda, Caveon

Presenters:
Sergio Araneda, Caveon
Gloria Ladson-Billings, University of Wisconsin-Madison
Stephen Sireci, University of Massachusetts Amherst
John Fremer, Caveon Test Security

FULL SCHEDULE

THURSDAY, APRIL 24

JOINT NCME/AERA-DIVISION D WELCOME RECEPTION

Thursday, April 24

6:30 PM TO 8:00 PM

Hilton Denver City Center Lower Level 2: Colorado Foyer

AWARDS CEREMONY AND MEETING

Friday, April 25: 7:30 AM to 8:30 AM

Hilton Denver City Center Lower Level 2: Colorado EF

PRESIDENTIAL ADDRESS

Chair: Amy Hendrickson, NCME Vice President & President-Elect

Speaker: Andrew Ho, NCME President

Friday, April 25: 8:30 AM to 9:30 AM

Hilton Denver City Center Lower Level 2: Colorado EF

089. NCME Career Award
Invited Session
9:45 AM – 11:15 AM
Hilton Denver City Center Lower Level 2: Colorado EF

Chair:
Daniel Bolt, University of Wisconsin – Madison

Presentation:
It's All About the Students
Terry Ackerman, University of Iowa

Discussants:
Ye Ma, Amazon Web Services
Devdass Sunnassee, University of North Carolina at Greensboro
Ronna Turner, University of Arkansas

090. AI Innovative Applications in Automated Scoring
Coordinated Paper Session
9:45 AM - 11:15 AM
Hilton Denver City Center Lower Level 2: Colorado AB

The use of generative AI (gen-AI) in automated scoring (AS) presents potential benefits and risks. This coordinated paper session explores innovative applications of AI to address problems that arise in applications of AS. For instance, gen-AI can enhance AS engine training efficiency via data supplementation, reducing the costs of model development and deployment. On the other hand, gen-AI tools can be misused by examinees (and in return testing agencies) to create (and detect) text submitted under the guise that it was written by the examinee. Integrating gen-AI into AS, requires careful consideration due to potential unintended consequences. Two papers in this session describe how gen-AI methods can supplement AS engine training datasets via data augmentation.

The remaining three papers focus on detecting bad-faith responses in AS: two of these apply gen-AI methods to detect AI-generated responses, while the third evaluates a process for creating AS models to detect the use of templated written responses. Hong Jiao, Professor at the University of Maryland, will be the discussant for the session.

Session Organizer:
Edward Wolfe, Pearson

Chair:
Edward Wolfe, Pearson

Presentations:
Cyborg Data: Merging Human with AI Generated Training Data
Kai North, George Mason University; Christopher Ormerod

Calibrating Generative AI Essay Complexity to Student Essay Complexity for Data Augmentation
Justin Barber, Pearson; Edward Wolfe, Pearson

Detecting AI-Generated Argumentative Essays
Cihan Demir, Washington State University; Amy Burkhardt, Cambium Assessment; Susan Lottridge, Cambium Assessment; Alexander Kwako, Cambium Assessment

Generative AI Detection—A Comparison of Multiple Methods
Mike Ritthaler, Pearson; Justin Barber, Pearson; Edward Wolfe, Pearson

Detection of Inauthentic, Templated Essay Responses in English Language Assessments
Yashad Samant, Pearson; Lee Becker, Pearson; Scott Hellman, Pearson; Bradley Behan, Pearson; Sarah Hughes; Joshua Southerland, Pearson

Discussant:
Hong Jiao, University of Maryland

091. Evaluating the Psychometric Impacts of Cultural Representation in Item Contexts

Coordinated Paper Session
9:45 AM - 11:15 AM
Hilton Denver City Center Lower Level 2: Denver 5-6

Amid growing awareness that social, cultural, and linguistic experiences affect examinees' understanding of items and engagement with an assessment, there is renewed interest in the ways that item contexts—particularly cultural contexts—affect examinees' interactions with items and ultimately how they perform on the items. For large-scale assessments, it is important to understand how cultural contexts might impact the psychometric properties of items and therefore score comparability, particularly when different examinees are administered different sets of items. This session highlights quantitative, psychometric research that attempts address questions related to the psychometric comparability of item variants (e.g., with and without contexts representing different cultural groups); the effects of cultural item contexts on item length, difficulty, and duration; the relative performance of examinee groups on items with cultural contexts, and whether there is formative value in introducing cultural or ethical contexts in test content. The answers to these questions will help support appropriate score interpretations and uses for assessment programs working to enhance validity and equity.

Session Organizer:
Jeffrey Steedle

Chair:
Jeffrey Steedle

Presentations:

Findings from Culturally Responsive Assessments Comprising Original and Adapted NAEP Items
Sandip Sinharay, ETS; Matthew Johnson, ETS Research Institute

Psychometric Impacts of Enhancing Cultural Representation in Math Items
Montserrat Valdivia Medinaceli, Curriculum Associates; Jeffrey Steedle

Antibias Content Creation in Medical Assessments
Ye Tong, National Board of Medical Examiners; Leslie Keng, National Board of Medical Examiners; Kristina Oberle, National Board of Medical Examiners; Amy Morales, National Board of Medical Examiners

Psychometric Evaluation of Culturally Relevant Items in a College Admissions Test
Cristina Anguiano-Carrasco, ACT; Jill McVey, ACT, Inc.; Jeffrey Steedle

Discussant:
Joseph Rios, Center for Measurement Justice

092. Revolutionizing Big Data with Advanced AI and Machine Learning (SIGIMIE-Organized Session)

Coordinated Paper Session
9:45 AM - 11:15 AM
Hilton Denver City Center Lower Level 1: Penrose 2

The intricate nature of big data poses challenges to traditional psychometric methods but opens up new opportunities for integrating data science to address these issues. The Big Data SIGIMIE session highlights three novel studies utilizing machine learning in diverse areas to handle big data and includes a panel discussion from academia and industry on better preparing for taming big data in practice. The presentations in the first section shed light on how to close the gaps between the urgent methodological demands of big data and the power of machine learning; demonstrate how to use machine learning methods to predict missing values in large-scale assessments; and illustrate how advanced techniques, such as large language models, can automatically analyze student dialogue as a form of formative assessment. These presentations will introduce a further panel discussion in the second section on how to transform big data with advanced AI and machine learning techniques in real practice. The SIGIMIE invites both experts from assessment frontiers and scholars in the academic research field to discuss the emerging needs for consolidating techniques for big data and preparing for future developments.

Session Organizer:
Qiwei He, Georgetown University

Chairs:
Qiwei He, Georgetown University
Tiago Caliço, Organization for Economic Cooperation and Development

Presentations:

MxML Phase 3: Proposing Changes to Measurement Standards for Machine Learning Enhanced
Yi Zheng, Arizona State University; Steven Nydick, Duolingo; Sijia Huang, Indiana University Bloomington

Predicting Missing Response with Machine Learning and LLM using Process Data
Qiwei He, Georgetown University

A First Attempt at Knowledge Tracing in Tutor-Student Dialogues
Andrew Lan, University of Massachusetts, Amherst

Discussants:
Alina von Davier, Duolingo
Susan Lottridge, Cambium Assessment
Peter Halpin, UNC-Chapel Hill

093. Comparing Approaches to Reduce Vulnerabilities and Improve Performance of Large Language Models in Assessment

Individual Paper Session
9:45 AM - 11:15 AM
Hilton Denver City Center Lower Level 2: Denver 3

Chair:
Ting Zhang, American Institutes for Research

Presentations:

Evaluating Annotation Performance of GPT-4o Mini, RoBERTa, and Human Annotators
Chi Chang, Michigan State University; Vipul Jain, Michigan State University
This study compares the performance of GPT-4o mini with BERT-based models and human annotators in handling annotations in medical video transcripts. Results demonstrate that GPT-4o mini consistently outperforms both BERT models and human evaluators in terms of accuracy, precision, recall, and F1 scores, improving clinical skill assessment processes.

Augmenting LLMs with Knowledge Seeds for Clinical Question Generation
Lingchen Kong, University of Florida; William Muntean, National Council of State Boards of Nursing; Joe Betts, NCSBN
This study proposes and evaluates a novel LLMs pipeline for generating high-stakes clinical assessment items. The LLMs-generated Multiple Choice (MC) items demonstrated comparable option-level ecological validity and distractor plausibility to experts-created items, with the pipeline facilitating item development and accelerating the refresh cycle to maintain item bank integrity.

Investigating Appropriate Uses for Automated Item Generation Content
Fen Fan, National Board of Medical Examiners; Jerusha Henderek, NBME
This study investigated the appropriate uses of automated item generation (AIG) by using subject matter expert judgments on item pairs selected with NLP methods. The findings will provide guidance to organizations trying to filter through the volume of (AIG) items from their AIG models and shed

Chatbot Performance as a Second-Attempt Examinee on a Medical Board Exam
Kelsey Nason, James Madison University; Carol Barry, American Board of Surgery; Caroline Prendergast; Andrew Jones, American Board of Surgery
This study mimicked the examinee's experience of using an LLM-based chatbot to complete a medical board continuous certification assessment. Results showed that the chatbot would pass the exam every time with a second attempt. Chatbot consistency and item characteristics were also explored. Limitations and future directions are discussed.

Robust Short Answer Scoring
Sahar Yarmohammadtoosky, Kennesaw State University; Yiyun Zhou; Peter Baldwin, National Board of Medical Examiners; Victoria Yaneva, National Board of Medical Examiners; Saed Rezayi, NBME; Polina Harik, NBME; Brian Clauser
This study explores vulnerabilities in transformer-based grading systems for medical education, focusing on adversarial gaming strategies. By identifying key weaknesses, implementing adversarial training, and utilizing ensemble techniques and large language models like ChatGPT-4, the research enhances robustness and highlights the need for improvements to ensure reliability in AI-driven education tools.

Discussant:
Ting Wang

094. Computerized Adaptive Testing

Individual Paper Session
9:45 AM - 11:15 AM
Hilton Denver City Center Lower Level 2: Colorado IJ

Chair:
Andreas Frey, Goethe University Frankfurt

Presentations:

Using Multiple Maximum Exposure Rates in Computerized Adaptive Testing
Kylie Gorney, Michigan State University; Mark Reckase, Psychometric Solutions
We propose an item exposure control method that involves the use of multiple maximum exposure rates. The new method is shown to improve both item pool utilization and measurement accuracy simultaneously.

IRT Residual-Based Approach to Evaluating Item Parameter Drift in CAT
Hwanggyu Lim, Inha University; Kyung (Chris) Han, GMAC; Yongsang Lee, Inha University
This study introduces a residual-based item parameter drift (RIPD) detection framework for evaluating IPD in computerized adaptive testing (CAT). By adapting the residual-based DIF framework, RIPD addresses the limitations of traditional IPD detection methods. A simulation study demonstrates its practicality, providing a new tool for continuous IPD monitoring in CAT.

Modified Fade-Away Method for Conditional Item Exposure Control in CAT
Ye Yuan, GMAC; Sung-Hyuck Lee, GMAC; Kyung (Chris) Han, GMAC
The purpose of the study is to introduce a new variation of the Fade-away (FA) method to control conditional item exposure rate in the computerized adaptive testing (CAT). The study presents FA and Local FA methods and compares existing exposure control techniques like Randomesque method to evaluate their effectiveness.

Evaluating IPD Detection Strategies in MST Contexts
Meereem Kim, Korea Institute for Curriculum and Evaluation; Hwanggyu Lim, Inha University; Doyoung Park, Korea Institute for Curriculum and Evaluation
This study evaluated item parameter drift (IPD) detection methods in multistage adaptive testing (MST) contexts. The research compared IRT-based approaches. Findings highlighted the pseudo-count D□ method's effectiveness when few items exhibited IPD and revealed challenges when many items showed IPD in MSTs with a small total number of items.

Recalibrating CAT Items: Expected Common Item Overlap for Nonequivalent Groups
Jonathan Weeks, American Institutes for Research
This is a simulation study that examines the sufficiency of common item data under item adaptive designs when the items are recalibrated for nonequivalent groups. A simple formula was derived that provides a close approximation to the expected proportion of common items under various design conditions.

Discussant:
Joshua Goodman, National Commission on Certification of Physician Assistants

194. Innovation Demonstration 2
Electronic Board Session
9:45 AM – 11:15 AM
Hilton Denver City Center Lower Level 2: Colorado GH

Presentations:

1. **Innovations in assessment: Interactive Storybook Platform for Equitable Assessment for Knowledge (ISPEAK)**
   Alisha Wackerle-Hollman, University of Minnesota; Rebecca Nathan, Aviellah Curriculum Consulting; Lillian Duran, University of Oregon
   This presentation introduces the Interactive Storybook for the Equitable Assessment of Knowledge; an app-based preschool assessment in English and Spanish that used a justice-oriented antiracist validation (JAV) approach for design and testing. Participants will learn about the design process, view a demonstration and explore practical application of the JAV framework.

2. **Addressing Recruitment Issues in Educational Measurement through an Innovative Internship Program**
   Brian Leventhal, James Madison University
   The field of educational measurement faces a significant recruitment challenge. During this demonstration I will discuss a 7-week annual summer internship program that has evidenced promise in recruiting to the field. Internship educational programming, lessons learned, and scalability will be discussed.

3. **Showcasing a STEM Conversation-based Assessment with AI Agents**
   Jessica Andrews-Todd, ETS; Edith Graf, ETS; Carol Forsyth, ETS; Dante Cisterna, ETS; Yi Song, ETS; Jeremy Lee, ETS
   This paper presents an innovative STEM assessment that incorporates AI-based conversational agents to generate information about students' understanding of weather-related phenomena and scientific practices when collecting and analyzing data. This task will be presented as a hands-on activity, with attendees engaging with the content and conversing with the AI agents.

4. **Hybrid Approach for Understanding Educational Surveys Using Generative Model and Text Encoders**
   Sungjin Nam, ACT, Inc
   This work introduces a hybrid method, combining sophisticated language understanding skills of generative models and light-weight text encoders, to improve the comprehensibility and interactivity of clustering analysis with open-ended educational survey responses in scale.

5. **Digital Learning Platform Research Catalog: Improving Access to Validated Process Data**
   John Whitmer, Federation of American Scientists; Warren Li, Associate Research Scientist
   Private philanthropies and federal agencies have recently funded digital learning platforms to create validated research data. This demonstration will present a free catalog for platforms to describe available datasets and experimental opportunities, to disseminate grants to use these platforms, and for researchers to match their interests to these opportunities.

6. **A Measurement Decision Theory (MDT) R Shiny Application**
   John Denbleyker
   An R Shiny web application intended to promote the use of measurement decision theory (MDT) for semi-formative assessments (e.g., quizzes, end-of-unit knowledge checks) is illustrated where classification and/or evaluating mastery status is the desired reporting outcome. Various simulations comparing MDT and IRT classification accuracy are included in the application.

7. **Mathematics learning environments and COVID-19: Relationship across locale and key demographic groups**
   Paulina Grekov, University of Wisconsin-Madison; Brittany Boyd, American Institutes for Research; Alvaro Cruz, American Institutes for Research; Cadelle Hemphill, American Institutes for Research
   A current area of interest is the extent to which student learning environment (SLE) shifted between 2019 and 2022 with respect to COVID-19. We introduce a Shiny application providing both graphical and numerical depictions of changes to SLE across the United States for a variety of student groups.

8. **Evaluating Climate Literacy: A Socioculturally Responsive Assessment Strategy**
   Rebecca Young, Cognia; Mary-Alice Corliss, Cognia; Mary Kate Clausen, Cognia
   Culturally responsive assessments for the future must prioritize student agency to confront the myriad social and environmental injustices climate change exacerbates. We advocate for and share an interdisciplinary assessment model that helps bridge the current gap between progress on compulsory K-12 climate instruction and authentic evaluations of students' climate literacy.

9. **Modeling Attribute Relationships in Diagnostic Models with the R Package measr**
   Jake Thompson
   Diagnostic models are a powerful tool for understanding the relationship between categorical latent attributes. However, these methods are not often used in applied research due to in part to limited and inaccessible software. In this presentation we describe how free software, measr, can easily estimate and evaluate diagnostic models.

095. Through-Year/ Interim Assessment Designs and Analyses

Individual Paper Session
9:45 AM - 11:15 AM
Hilton Denver City Center Lower Level 2: Colorado CD

Chair:
Luciana Cancado, Curriculum Associates

Presentations:

Maximizing assessment information after instruction: using responses, time, and multiple attempts
Yikai Lu, University of Notre Dame; Erin Banjanovic
The context of personalized instruction differs from traditional assessments. If the instruction is effective, items are usually answered correctly, which limits their item information. This study explored IRT models that included additional sources of information: responses, time, and attempts. The results showed most models led to decreases in measurement error.

Multistage Testing Efficiency in Through-Year Assessments under Simulated Routing Error
Sangdon Lim, Cambium Assessment
Through-year assessments can utilize a multistage design over an academic year. One practical concern of using multistage design is the impact of incorrect routing on assessment outcomes, which may be more concerning when there are more stages. Operational configurations were used to simulate the impact of a worst-case routing error.

Evaluating the Texas Through-Year Assessment Pilot Participation Efficacy
Yuanyuan McBride, Pearson; Michael Chajewski, Pearson
This study evaluated TTAP participation's efficacy by comparing matched TTAP and non-TTAP examinees across multiple STAAR subjects. Coarsened Exact Matching (CEM), weighting strategy, and regression analyses were used. Results showed significant impacts of TTAP in select subjects, though overall effect sizes were negligible, indicating limited practical significance.

A Comparison Between Full vs Partial Blueprints in Through-Year Assessment Combined Scoring
Maryam Pezeshki, Cambium Assessment Inc.; Shuqin Tao, Cambium Assessment Inc.; Elizabeth Wei, Cambium Assessment Inc.; Stephan Ahadi, Cambium Assessment Inc.
One of the challenges involved in planning through-year-assessment (TYA) designs is if/how the content needs to be segmented. States adopting a TYA design use either full blueprint or partial blueprint for interim assessments. This study compares the two strategies for meaningful end-of-year combined score reporting. Implications are included for educators.

Discussant:
Sonya Powers

096. Content and Internal Structure Validity Evidence

Individual Paper Session
9:45 AM - 11:15 AM
Hilton Denver City Center Lower Level 2: Denver 1-2

Chair:
Yangmeng Xu, University of Alabama

Presentations:

Enhancing interview protocols: topic modeling as a content validity technique
Constanza Mardones-Segovia, University of Georgia; Tamlyn Lahoud, University of Georgia; Cheng Tang, University of Georgia; Yaxuan Yang, University of Georgia Shiyu Wang, University of Georgia; Allan Cohen, University of Georgia; Kun Wang, Rethink Learning Labs; Chandra Orrill, Rethink Learning Labs; Rachel Brown, The Pennsylvania State University

This empirical study proposes a data-driven framework for evaluating content validity in interview data using Natural Language Processing and a Structural Topic Model (STM). Preliminary results suggested that STM is a promising quantitative technique for content validity in unstructured data, reducing the time and effort needed to refine interview protocols.

Assessing College Readiness: Internal Structure Validity Evidence for Multistage Diagnostic Reading Assessment

Oxana Rosca, University at Albany, SUNY; Angela Lui, City University of New York; Kimberly Colvin, University at Albany, SUNY; Heidi Andrade, University of Albany; Jason Bryer, City University of New York

This study aimed to gather validity evidence regarding internal structure for a free, open-source, untimed testlet-based diagnostic reading assessment that offers students immediate personalized feedback. Two-dimensional testlet-based IRT model fits (bifactor 1PL and 2PL) were assessed using ANOVA nested models' comparisons, local dependence residuals, item parameters, and standard errors.

Assessing A Learning Progression in Science: A Unidimensional versus A Three-Dimensional Construct

Mingfeng Xue, University of California, Berkeley; Mark Wilson, University of California, Berkeley; Linda Morell, University of California, Berkeley

We develop a three-dimensional assessment for NGSS learning progressions, focusing on argumentation (ARG), pattern recognition (PAT), and structure of matter (SOM). The results show that the three-dimensional model fits better than a unidimensional one, revealing the complex relationships between the constructs and providing valuable insights for improving science education curricula.

Student Self-Expectations of The First-Year College Experience (The SE-FYE Scale): Scale Validation

Min Xiao, University of Kentucky

This study aims to develop, validate, and refine the SE-FYE scale by applying the Rasch measurement analysis. The scale was refined to 22 items across five constructs: general first-year expectations, academic readiness, engagement, personal development, and career preparation supported by reasonable unidimensionality, fit, separation, reliability, and category functionality.

Discussant:
Chad Buckendahl, ACS Ventures, LLC

097. How Can We Balance Assessment Systems to Enable Teacher Collaborative Learning?

Organized Discussion
9:45 AM - 11:15 AM
Hilton Denver City Center Lower Level 1: Penrose 1

A primary purpose of balanced assessment systems is to improve classroom instruction through school-based professional learning for teachers. Yet research on teacher learning and assessment systems rarely speak to one another. In this interdisciplinary discussion, we examine how balanced assessment systems can be organized to enable productive teacher collaborations in order to realize and sustain ambitious and equitable teaching and learning, particularly for minoritized children and youth in low-income communities. We review the criteria of balanced assessment systems and grapple with why, after a generation of advocacy, they remain so rare in U.S. education. Additionally, we identify the structural conditions of—and critical processes involved in—teacher collaborations that produce meaningful changes in teaching, improved student achievement, and greater teacher efficacy and fulfillment. We discuss possibilities and priorities for interdisciplinary partnerships to design assessment systems that enable teacher collaborative learning in a new era of policymaking in U.S. public education. We will analyze how current assessment systems both afford and constrain teacher collaborative learning, examine the process of teacher reasoning and decision-making about their instruction using assessment information, and discuss possibilities for assessments of and for teaching.

Chair:
Scott Marion, Center for Assessment

Presenters:
Courtney Bell
Bryant Jensen, Brigham Young University
Adam Lefstein, The Hebrew University of Jerusalem
Lorrie Shepard
Sarah Woulfin, The University of Texas at Austin

098. NCME Task Force on the Role of Educational Measurement and Civil Rights (NCME Committee Organized Session)

Coordinated Paper Session
11:30 AM - 1:00 PM
Hilton Denver City Center Lower Level 2: Colorado EF

Education, directed to the full development of the human person, is recognized as a fundamental human right (United Nations, 1949). In the United States, education rights are enshrined in constitutional protections and state laws. Persons in the United States have rights to education that may not be abridged on the basis of gender, race, disability, language, and other protected categories. In 2023, the Board of Directors of the National Council on Measurement in Education (NCME) reflected on the roles, positive and negative, that educational tests and assessments have played in relation to civil rights. Recognizing the complexity of the issues, the criticality of accurate information, and the timeliness of a substantive contribution to public discourse, a Task Force on the Role of Educational Measurement in the Protection and Advancement of Civil Rights was conceived, charged with creating clear and accurate information regarding educational tests and assessments that is relevant to the protection and advancement of the civil rights of all to education. In this session, task force members will present their work to the NCME community, initiating a solicitation of public comment on the Task Force work to date.

Session Organizer:
Britte Cheng, Menlo Education Research

Chair:
Britte Cheng, Menlo Education Research

Presentations:

NCME Presidential Task Force on Civil Rights Report: Purpose, Process, and Themes
Britte Cheng, Menlo Education Research; Michael Walker, HumRRO; Natalie Rambis, Menlo Education Research

Educational Measurement and Civil Rights: Key Concepts and Case History
Sarah Beach, University of Virginia; Magdalen Beiting-Parrish, Federation of American Scientists; Carlos Chavez; Anne Davidson, EdMetric; Jessica Hoskins

Civil Rights Protections Through Educational Measurement Standards
Anne Davidson, EdMetric; Sarah Beach, University of Virginia; Magdalen Beiting-Parrish, National Center for Education Statistics/ Federation of American Scientists; Jessica Hoskins; Carlos Chavez

The Landscape of Educational Measurement Programs Through the Lens of Civil Rights
Juan D'Brot, Center for Assessment; Carlos Chavez; Regina Lewis, Maine Department of Education; Catherina Villafuerte, University of Connecticut; Michael Walker, HumRRO

Civil Rights Considerations in the Development of Legislation, Policy, and Practice
Howard Everson; Maria Oliveri, Purdue University; Mya Poe, Northeastern University; Michael Russell, Boston College; Lauren White

Where do we go from here?
Michael Russell, Boston College; Magdalen Beiting-Parrish, National Center for Education Statistics/ Federation of American Scientists; Britte Haugan Cheng, Menlo Education Research

Discussant:
Michael Walker, HumRRO

099. Dive into Item and Form Development Challenges in Large-Scale Assessment

Coordinated Paper Session
11:30 AM - 1:00 PM
Hilton Denver City Center Lower Level 2: Colorado CD

Despite the widespread usage of test scores in various evaluations and decision-making, item and form development has yet to be the focus of research in the measurement community. This coordinated session aims to address this gap. The first two papers will discuss item-level investigations into development. The third paper evaluates the usage of various online tools to better inform item development. The last two papers contribute to form development, particularly in multistage computerized adaptive testing (MST). Attendees will enhance their understanding of item and form development challenges and contemporary thinking on possible solutions.

Session Organizer:
Ji Zeng

Chair:
Ji Zeng

Presentations:
Using RTD to Address Two Major Obstacles to Developing Valid Items
Marjorie Wine, Accessible Teaching, Learning and Assessment Systems (ATLAS), University of Kansas; Alexander Hoffman, AleDev Research & Consulting

Critical Examination of Single-item Self-perception Constructs in a Large-scale Assessment
Serife Bilgic-Erdem, William & Mary; Onur Ramazan, Washington State University

Toward Better Digital Tool Designs to Assist Students on Math Assessments
Hongwen Guo, ETS; Matthew Johnson, ETS Research Institute; Luis Saldivia, ETS; Michelle Worthington, ETS

Developing and Evaluating Psychometric Criteria for Multistage Tests
Louis Roussos, Cognia; Qi Qin, Cognia; Liuhan Cai, Accenture; Katie Schmidt, College Board

Effects of the Highly Adaptive Testing Design for PISA on the Precision of Plausible Value-Based Population Estimates
Andreas Frey, Goethe University Frankfurt; Christoph König, Goethe University Frankfurt, Germany; Aron Fink, Johann Wolfgang Goethe University

Discussant:
Andrew Middlestead, Michigan Department of Education

100. Approaches to Optimizing a Personalized Learning System

Coordinated Paper Session
11:30 AM - 1:00 PM
Hilton Denver City Center Lower Level 2: Colorado IJ

This set of research studies illustrates approaches to refining a personalized learning system wherein students work at their own pace through a set of learning materials, practice items, and mastery quizzes. The first study tackles a complicated calibration design where the concept of common items and common persons is stretched. Because items are taken across an academic year, various time

interval durations are explored to see how short of a time interval is needed to obtain reasonably accurate item difficulty estimates. After placing items onto a common scale, study two investigates whether the articulation of difficulty across the learning progression is consistent with expectation. Item pool difficulty targets are established using methods more commonly applied in standard setting. Difficulty targets can be used to cull items from the pools or, as illustrated in study three, to refine item selection methods that leverage the entire item pool while controlling for overall form difficulty. To maintain an adequate number of items that are of appropriate content and difficulty, study four aims to evaluate the feasibility of using GPT for automated item generation. By leveraging psychometric and AI approaches, user experience and data-driven insights about student learning will be improved.

Session Organizer:
Sonya Powers

Chair:
Sonya Powers

Presentations:
Calibrating Items across a Learning Progression
Jinah Choi, Edmentum

Developing Difficulty Targets for Items in a Learning Progression
Yi He, Edmentum; Jinah Choi, Edmentum

Optimizing Test Assembly for Mastery Quizzes
Ziyu Fan, The University of Texas at Austin

Generating Items with Content and Difficulty Specifications using Large Language Models
Pooya Razavi, Edmentum

Discussant:
Tracey Hembry, Pearson

101. Innovations in Automated Scoring Modeling

Coordinated Paper Session
11:30 AM - 1:00 PM
Hilton Denver City Center Lower Level 2: Colorado AB

Automated scoring (AS) of constructed responses relies heavily on machine learning technologies to identify optimal models for predicting scores that would be assigned by human raters. This coordinated paper session features five presentations that focus on innovative methodologies for creating, optimizing, and evaluating AS models. The first paper provides a framework for documenting the validity of score interpretations associated with feature-based and generative AI automated scoring models. The second paper evaluates the promise of ordinal regression as an approach to creating AS models. The third paper explores the use of Bayesian decision theory as a means of creating AS models that optimize weighted coefficient kappa. The fourth paper examines the relationship between writing process data as a potential predictor of measures of writing quality. The fifth paper proposes a Bayesian approach for evaluating the fairness of automated scores. Sue Lottridge of Cambium will serve as the session discussant.

Session Organizer:
Edward Wolfe, Pearson

Chair:
Edward Wolfe, Pearson

Presentations:
Validity Evidence for Use and Interpretation of Scores from Generative AI
Jodi Casabianca, ETS; Vladimir Zubenko, ETS; Alper Naim, ETS; Daniel McCaffrey, ETS

Evaluating Hybrid Ordinal Regression Models for Enhanced Automated Essay Scoring Accuracy
Martha Bellows, Pearson; Justin Barber, Pearson; Edward Wolfe, Pearson

Training Automated Scoring Engines by Optimizing for Kappa Using Decision Theoretic Approaches
Matthew Johnson, ETS Research Institute; Michael Fauss, ETS

Enhancing Automated Scoring with Writing Process Data
Hong Jiao, University of Maryland; Ming Li, University of Maryland; Tianyi Zhou, University of Maryland; Chandramani Lnu, University of Maryland; Nan Zhang, University of Maryland

An Empirical Bayesian Approach for Testing the Fairness of Automated Scoring
Sunbeom Kwon, University of Illinois, Urbana-Champaign; Daniel McCaffrey, ETS; Paul Jewsbury, ETS; Jodi Casabianca, ETS

Discussant:
Susan Lottridge, Cambium Assessment

102. Improving Score Reporting of Student Group Results to Better Support Educational Equity

Coordinated Paper Session
11:30 AM - 1:00 PM
Hilton Denver City Center Lower Level 1: Penrose 1

This session on large-scale assessment score reporting explores innovations to improve reporting and increase its potential use to address inequities in educational opportunities. Traditional approaches to score reporting, such as focusing on “achievement gaps” in mean racial group scores, have been found to unintentionally lead to deficit thinking and misrepresentation of students from lower-performing groups. That is, the blame for lower performance is placed on student-group identity instead of systems or resource inequities. Our presentations detail specific approaches to address this issue through changes to data visualizations and other score report features. Presented work includes (1) research on a collaborative score report codesign process involving all session presenters, (2) data visualization experiments on attribution biases, (3) a framework for data collection, accountability, and reporting for English learners, and (4) an overview of the process behind and impacts of a state’s practice of combining student groups in reporting. Attendees will learn about potential approaches to improve data visualizations and reporting to reduce biased interpretations and better inform productive approaches to address group score differences.

Session Organizer:
Marisol Kevelson, ETS

Presentations:
Using Codesign to Develop Equity-focused, Anti-deficit Data Displays
Vinetha Belur, ETS; Katherine Furgol Castellano, ETS; Marisol Kevelson, ETS; Emily Kerzabi, ETS; Eli Holder; Leslie Villegas; Renee Savoie; Jennifer Cain, Minnesota Department of Education

Painting a Complete Picture of Students Identified as English Learners Through Better Data Collection and Reporting
Leslie Villegas

Maximizing Impact: The Benefits of Supergroups in Accountability Reporting
Renee Savoie

Can Bar Charts Be Racist?
Eli Holder

Discussant:
Priya Kannan, WestEd

103. Graduate Student eBoard Session 3

Graduate Student Electronic Board Session
11:30 AM - 1:00 PM
Hilton Denver City Center Lower Level 2: Colorado GH

Presentations:

1. **Exploring Automated Scoring Using Topics and Predefined Features in Hybrid Attention-Based Model**
   Cheng Tang, The University of Georgia; George Engelhard, University of Georgia; Jiawei Xiong, Curriculum Associates
   This research proposes a Bidirectional Long Short-Term Memory network with attention mechanism, topic modeling, and feature extraction to enhance automated scoring. Using the Automated Student Assessment Prize dataset, the model achieved an average Quadratic Weighted Kappa value of 0.74 across three prompts, demonstrating the potential for reliable automated educational assessments.

2. **Internal structure and invariance of a functional competency assessment for Indian youth**
   Ketan, University of Massachusetts, Amherst; Craig Wells, University of Massachusetts Amherst
   This study evaluates the internal structure and measurement invariance of the Beyond Basics assessment, which measures functional competencies in youth aged 14 to 18. Using confirmatory factor analysis and differential item functioning, we examine dimensionality and fairness across gender and education groups in a large-scale sample from rural India

3. **Does Varying CFA Fit Indices Cutoffs Impact Standard Errors of Rasch Parameters?**
   Pablo Vivas Corrales, University of Minnesota; Mohammed Abulela, University of Minnesota; Michael Rodriguez, University of Minnesota
   We examined the association between ordinal CFA fit indices and the Rasch model standard errors. Through Monte Carlo simulations, we observed that traditional cutoff criteria may not guarantee precision on Rash model parameters. Applied researchers should tailor fit criteria to align with the unique demands of their analysis.

4. **Inquiry-Based Teaching Profiles and Their Effect on Student Achievement**
   Zhijun Chen, Department of Education, University of Bath; Diego Carrasco, Centro de Medición MIDE UC, Pontificia Universidad Católica de Chile; Andres Sandoval Hernandez, Department of Education, University of Bath
   This study uses TIMSS 2019 to explore the identification of distinct teacher profiles in Arabic countries based on inquiry-based teaching practices. Using multigroup latent class analysis and ANOVA, the research examines the relationship between these profiles and student achievement in math and science, revealing significant educational implications for teacher development.

5. **Exploring Machine Learning Methods for Measuring Student Engagement in a Low-stakes Assessment**
   Jack Gilmore, James Madison University; Yu Bao, James Madison University
   This study explores the use of machine learning as an alternate method of identifying non-effortful responses in low-stakes assessments. By comparing supervised and unsupervised machine learning techniques to traditional approaches, this research aims to enhance the validity and reliability of low-stakes assessment scores in higher education.

6. **Assessing the Impact of Empirical Histogram RMSD Variants on DIF Detection**
   Preston Botter; Yuan-Ling Liaw, IEA Hamburg; Rolf Strietholt, IEA Hamburg; Dubravka Svetina Valdivia, Indiana University
   This study explores the impact of four root mean squared deviation (RMSD) variants on DIF detection in the 2023 CIL assessment. Empirical histogram (EH) methods reveal that weighting schemes influence DIF identification, highlighting the need for researchers to include EH variants in ILSA data sensitivity analyses.

7. **Predicting Reading Achievement in PIRLS 2021: A Comparative Analysis of Machine Learning Models**
   Fatih Ozkan
   This study evaluates the effectiveness of Gradient Boosting, Support Vector Regression, and Decision Tree Regressor in pre- dicting reading achievement using PIRLS 2021 data. Feature selection identified socioeconomic status and parental education as key predic- tors. Gradient Boosting outperformed other models, demonstrating superior predictive accuracy, with significant implications.
8. **Fairness of Text-To-Speech as Designated Support on Statewide Assessments: A DIF Analysis**
   Annette Vernon, University of Iowa; Catherine Welch
   This study applied DIF detection methods to examine fairness and comparability of a statewide reading comprehension assessment, treating students with access to text-to-speech technology as the focal group and those without as the reference group. Findings provide evidence for comparability and validity with implications for test developers and policy-makers.
9. **Remote Testing in a State Summative Program: Exploratory and Differential Item Analysis**
   Haoyang Yu, University of Kansas; Rebecca Freund, Pearson
   This study profiles remote test takers in a state summative assessment program. Item-level analyses include differential item functioning using Mantel-Haenszel and standardized mean difference methods, as well as differential response time statistics, to explore potential disparities between remote and on-site testers.
10. **Measurement invariance and factor structure—What if Configural Invariance is wrong?**
    Rujun Xu, University of Virginia; James Soland, University of Virginia
    This study investigated how well invariance testing with MGCFA performs in identifying non-invariance in factor structure between subgroups of a dataset. By generating various non-invariance patterns with Monte-Carlo simulation, we found that in most of the conditions, MGCFA can make mistakes in concluding configural invariance. Findings are discussed.
11. **DIF in PISA 2022: Comparison of Differential Item Functioning Detection Methods**
    Woori Song; Yejin Kim
    This study compares the Rasch model, logistic regression, Mixture IRT, and Rasch-tree method to detect Differential Item Functioning (DIF) in PISA 2022 mathematics data. Results show that Rasch-tree is more effective at identifying DIF, considering a broader range of individual characteristics compared to traditional methods.
12. **Impact of Linking Methods on Vertical Scaled Assessments**
    Stephanie LeRoy; Christine Demars, James Madison University
    This study examines the impact of different linking methods on vertical scaled assessments. Using simulations, direct and chain linking designs were crossed with mean-only and mean-sigma linking methods under the 3PL and Rasch models. Results indicated that mean-only and mean-sigma methods impacted cohort and growth differently.
13. **Regression Discontinuity Design Within a Latent Variable Framework: Accuracy of the Treatment Effect**
    Sarah Hammami; James Bovaird, University of Nebraska-Lincoln
    This study explores the multiplicative effect of a) loading magnitude, b) number of indicators, and c) sample size on the Type 1 error rate, relative bias, and power of the treatment effect in regression discontinuity designs when the outcome is framed within a structural equation modeling framework.
14. **Comparing Unidimensional Latent Score Metrics for Three Vertical Scaling Methods Under Multidimensionality**
    Lionel Meng, University of Wisconsin – Madison; Xiangyi Liao, University of Wisconsin Madison; Daniel Bolt, University of Wisconsin - Madison

The latent score metrics that result from 2PL, Bifactor, and Projection IRT vertical scaling methods in the presence of construct shift multidimensionality are compared. Score metrics are evaluated according to (1) the marginal IRT reliability of scores within grade, and (2) stability of growth metrics between grades.

15. **Exploring the Efficacy of Including Technology-Enhanced Items in Large Scale Assessments**
Ahmed Bediwy, The University of Iowa; Alexis Oakley; Anthony Fina, University of Iowa; Catherine Welch
In recent years, there has been a rise in conducting large-scale state assessments digitally, creating a demand for new item formats capable of evaluating content knowledge and essential skills more comprehensively. This study aims to assess the performance of TEIs compared to traditional MCIs in large-scale state assessments.

16. **Comparison of Different Machine Learning Methods in the Prediction of College Readiness**
Jeongmin Ji, University of Iowa; Catherine Welch; Stephen Dunbar, University of Iowa
With the increase in the use of artificial intelligence, prediction models based on machine learning methods can help accurately measure college readiness. Using five supervised learning methods, this study explores factors that influence college readiness and compares model performance to find the best model to predict college readiness.

17. **An Item-Level Investigation of the Impact of Kumon**
Lien Vu, University of Delaware; Sanford Student, University of Delaware; Richard Houang, Michigan State University
This proposal presents new information on the impact of Kumon on math competencies using item and overall level effects. It will also investigate the relationship of the item responses by strand (or groups of items) and the relative positions of Kumon students in their Kumon study (e.g. duration, level).

104. New Models and Methods for DCM

Individual Paper Session
11:30 AM - 1:00 PM
Hilton Denver City Center Lower Level 2: Denver 1-2

Chair:
Nikole Gregg, National Commission on Certification of Physician Assistants

Presentations:

A Bayesian Multinominal Model for Cognitive Diagnosis with Multiple-choice Items
Lawrence DeCarlo, Teachers College, Columbia University
Multiple-choice items in cognitive diagnosis can be analyzed using a general multinomial model with latent predictors, which includes, as special cases, the MC-DINA model (de la Torre, 2009) and the MC-SDT model (DeCarlo, 2021a). A Bayesian version of the multinomial model is shown to provide a bridge between the models.

A Comparison of Person-Fit Indices for Cognitive Diagnostic Modeling
Audrey Filonczuk; Ying Cheng, University of Notre Dame
We propose the SSR person-fit statistic for detecting aberrant responders in CDM, comparing it to the MSR (Yu & Cheng, 2019, 2020) and RCI (Cui & Li, 2015). Simulations were conducted to, first, derive empirical distributions of the residuals and, second, reveal superior detection capabilities of the SSR.

Rating Scale Diagnostic Classification Model with Dispersion Parameter
Alexandra Kislyonkova, University of California Merced; Ren Liu, University of California, Merced
The current study introduces a Diagnostic Classification Model (DCM) for rating-scale data, easing constraints of existing models while remaining theoretically sound. It is tested through simulation to verify parameter recovery and in an applied study to compare its performance with existing DCMs for rating scales.

A Multi-strategy Cognitive Diagnosis Model Incorporating Item Response Times and Fixation Counts
Junhuan Wei, Jiangxi Normal University; Dongbo Tu, Jiangxi Normal University; Yan Cai, Jiangxi Normal University
Based on strategy selection theories, this study develops a multi-strategy cognitive diagnosis framework integrating response accuracy (RA), response times (RTs), and fixation counts (FCs) to enhance diagnostic feedback on strategy selection and cognitive processes, improving individual diagnosis accuracy.

Discussant:
Yon Soo Suh, NWEA within HMH

105. DEI and Language Learning

Individual Paper Session
11:30 AM - 1:00 PM
Hilton Denver City Center Lower Level 2: Denver 4

Chair:
Kerry Englert, Seneca Consulting, LLC

Presentations:

Impact of Test Scoring Methods on Language Learners
Yifei Li; Tim Moses, Buros Center for Testing; Luz Bay, College Board; YoungKoung Kim, The College Board
Performance gaps for language learners persist despite decades of improvements in education and standardized testing. This study considers the role of test scoring methods. Real data and simulations show that performance gaps for language learners are larger for some scoring approaches, primarily due to group differences in item parameters.

Examining English Learners' Proficiency Before and After the Pandemic: Unpacking language domains
Narék Sahakyan, University of Wisconsin - Madison / WIDA
Leveraging multilevel models accounting for the clustering of millions of EL students across schools, districts, and states, and student-level factors such as "newcomer" or "long-term" status, age, ethnicity, race, etc., I present evidence of the pandemic's persistent impact and growing disparities in EL proficiency between Hispanic and non-Hispanic identified ELs.

Assessment of Calibration Methods for World Language Examinations
Yuan Ge, College Board; Tim Moses, Buros Center for Testing; YoungKoung Kim, The College Board
This study examines calibration and scoring methods for native speakers and language learners in large-scale world language exams. The researchers compared item calibrations using 1PL and 3PL IRT models for standard and total groups, aiming to improve calibration procedures and ensure fair assessments for all test takers.

Developed For and With Californian Children: A Reading Screener for Spanish-English Multilinguals
Julian Siebert, Stanford University; Mónica Zegers, UCSF; Lillian Duran, University of Oregon; Marilu Gorno Tempini, University of California, San Francisco
We developed a suite of tasks assessing Spanish (pre-)reading skills and, using a sample (N = 612) representative of California's multilingual population, built a universal screening tool that accurately identifies students struggling to learn to read in Spanish. We discuss the most predictive constructs and use cases for Spanish screening.

Discussant:
Cristina Anguiano-Carrasco, ACT

106. Scaling and Linking with IRT

Individual Paper Session
11:30 AM - 1:00 PM
Hilton Denver City Center Lower Level 2: Denver 5-6

Chair:
Seongeun Kim

Presentations:

Psychometric Properties of Nonlinearly Transformed Scale Scores from Rasch Estimates
Juyoung Jung, University of Iowa; Won-Chan Lee, University of Iowa
This study examines the psychometric properties of nonlinearly transformed scale scores derived from Rasch estimates on mixed-format tests. The delta method is used to approximate conditional standard error of measurement and reliability for the scale scores. Classification errors and related indices are estimated within the Rasch framework, employing various approaches.

Linking Scales Using the Measurement Alignment Method for the Simple-Structure MIRT Model
Seongeun Kim; Kyung Yong Kim, University of North Carolina Greensboro
This study extends a measurement alignment method, a viable alternative linking method, to the simple-structure MIRT model and compares its relative performance to three commonly used linking methods. The feasibility of the extended measurement alignment method is evaluated through a simulation study with various study conditions and real-data analysis.

Perspectives on Intervals in Scaling: Ability Measures vs. Test Performance Indicators
Tim Moses, Buros Center for Testing; Jinghua Liu, The National Board of Osteopathic Medical Examiners; Stuart Barnum, The National Board of Osteopathic Medical Examiners
This study considers approaches to establishing reporting scales with interval interpretations, from Rasch ability estimates and observed test scores. Scale scores are compared in empirical data. Evaluations consider inconsistencies when scales produced from ability estimates or observed scores are interpreted with respect to observed scores or ability estimates.

Curse of compounding error: Exploring linking constant variation in small samples
Brandon LeBeau, WestEd; Meredith Langi
When conducting item response theory (IRT) linking, the IRT estimates are treated as fixed. Then, the linking constants are estimated to rescale parameters across groups. In small samples, the error in the IRT estimates will impact the estimation of the linking constants, and we estimate the variation in this error.

Discussant:
Stella Kim, University of North Carolina Charlotte

107. Federal Policy Facilitators and Barriers for Performance Based Assessment Systems: National Perspectives on a Case Study

Organized Discussion
11:30 AM - 1:00 PM
Hilton Denver City Center Lower Level 1: Penrose 2

This session brings together researchers and practitioners with expertise in psychometrics, content alignment, policy, and assessment development to discuss federal policy facilitators and barriers for innovative assessments within state accountability systems. Performance-based measures are generally considered to provide assessment opportunities that are more authentic, motivating, and instructionally and culturally relevant than more standardized formats. Under ESSA, the Innovative Assessment Demonstration Authority (IADA) is included to promote the development and use of high-quality innovative assessments, including tests consisting of fully performance-based measures. However, the US Department of Education (ED) has struggled to enlist and maintain

state participation. State feedback to ED emphasizes concerns related to meeting comparability and alignment requirements as stated in federal statute (The Education Trust, 2023). To investigate the compatibility of innovative assessments with current interpretations of federal requirements associated with comparability and alignment, our research team evaluated the comparability and alignment a performance assessment program that was piloted by four districts within a state in 2023-2024. Results from the analyses of the case study will be shared, along with an interactive discussion on the implications for policy. Presenters will discuss actionable recommendations for improving the compatibility of federal policy and innovative measures within state accountability systems.

Chair:
Susan Lyons, Lyons Assessment Consulting

Moderator:
Michelle Odemwingie, Achievement Network

Presenters:
Sara Christopherson, Wisconsin Center for Education Products and Services (WCEPS)
Samuel Ihlenfeldt, University of Minnesota
Julianna Brown, KnowledgeWorks
Michelle Odemwingie, Achievement Network

108. Computational Psychometrics in the Wild: Scalable Procedures for Maintaining Test Scores

Coordinated Paper Session
1:30 PM - 3:00 PM
Hilton Denver City Center Lower Level 2: Denver 1-2

Psychometricians have long advocated including sophisticated computer science algorithms (such as those from machine learning, artificial intelligence, and data mining) in psychometric applications to improve the efficiency of measurement. However, these applications have been bound by the immediate requirements of test scores and the limits of technology. The debate of how to include modern computational methods in the service of measuring psychological traits (so-called “Computational Psychometrics”) has recently been upended by theoretical and practical advancements in artificial intelligence (AI) and machine learning (ML). Operational testing programs might wish to consider adopting a computational mindset but lack practical examples of how one can integrate Computational Psychometrics into the test maintenance lifecycle. This session demonstrates how computational methods are currently used to generate and maintain defensible scores in practice. The corresponding presentations include Computational Psychometrics in the service of item development (using AI), item calibration using features to better predict item properties, adaptively administering items and scoring persons with item parameter uncertainty, and scaling test scores to maintain both subscore and overall score distributions. All of these procedures use modern computational methods to solve challenges in maintaining test scores and scaling item development processes given limited resources and operational requirements.

Session Organizer:
Steven Nydick, Duolingo

Chair:
Steven Nydick, Duolingo

Presentations:
Ensuring Content and IRT-Parameter Distributions in Item Generation for Adaptive Tests
Andrew Runge, Duolingo; Phoebe Mulcaire, Duolingo; Kevin Yancey, Duolingo; James Sharpnack, Duolingo; Kai-Ling Lo, Duolingo; Geoff LaFlair, Duolingo

A Scalable Item Parameter Estimation Algorithm for Large and Highly Sparse Data
Steven Nydick, Duolingo; J.R. Lockwood, Duolingo

A Thompson Sampling Approach to IRT-based Computerized Adaptive Tests
James Sharpnack, Duolingo; Kevin Hao, Duolingo; J.R. Lockwood, Duolingo; Steven Nydick, Duolingo; Alina von Davier, Duolingo

Establishing Score Reporting Scales Under Constraints: Subscores for the Duolingo English Test
Manqian Liao, Duolingo; Steven Nydick, Duolingo; J.R Lockwood, Duolingo

Operationalizing a Personalized Ensemble Test (PET)
Alina von Davier, Duolingo; Kevin Yancey, Duolingo

Discussant:
Stephen Sireci, University of Massachusetts Amherst

109. Competency-Based Education: Theory, Framework, and Practice

Coordinated Paper Session
1:30 PM - 3:00 PM
Hilton Denver City Center Lower Level 2: Colorado CD

Competency-based education (CBE) has drawn incredible attention in the United States as both K-12 and higher education call for a heightened focus on mastery and skills. An effective education system needs to recognize and promote learning, instead of constraining learning through fixed seat time and credit hours. In this system, students are empowered daily to own decisions about their learning experience and demonstrate progress based on evidence of mastery of knowledge and skills. This system also recognizes and supports multiple pathways of student progress, and employs strategies to ensure equity for students from diverse backgrounds. This coordinated session aims to shed light on the most salient aspects of competence-based education, including four presentations that discuss (a) key elements to facilitate system-wide changes towards CBE, (b) learning progressions developed to describe critical skills development in CBE, (c) innovative school models in implementing CBE and aligning with skills frameworks, and (d) how non-credit certification can be granted to expand CBE in higher education.

The four presentations jointly provide a unique opportunity for the audience to gain insights into CBE, connecting researchers and practitioners.

Session Organizer:
Ou Lydia Liu, ETS

Chair:
Ou Lydia Liu, ETS

Presentations:
Towards a Research and Development Agenda for Competency-Based Assessment
David Sherer, Carnegie Foundation; Norma Ming, Carnegie Foundation for the Advancement of Teaching

Learning Progressions to Support Measuring and Developing Essential Skills
Lei Liu, ETS; Teresa Ober, ETS; Kevin Williams, ETS; Guangming Ling, ETS; Karyssa Courey, Rice University

CBE Initiatives and Alignment with an Ecosystem to Support Students' Skill Proficiency
Teresa Ober, ETS; Lei Liu, ETS; David Nitkin, Transcend Education; Ou Lydia Liu, ETS

Providing Credit for Non-Credit Certificates, One Institution's Journey
Kristen Hodge-Clark, Georgetown University School of Continuing Studies; Jeffrey Warner, Georgetown University School of Continuing Studies

Discussant:
Kristina Zeiser, AIR

110. It's Time To Evaluate State Accountability Systems

Coordinated Paper Session
1:30 PM - 3:00 PM
Hilton Denver City Center Lower Level 2: Denver 3

Too often, accountability systems are designed and implemented without an evaluation and continuous improvement plan. All state accountability systems that meet the Every Student Succeeds Act (ESSA) requirements have their plans reviewed and approved by the U.S. Department of Education. States must also report their annual results to the U.S. Department of Education. However, these systems rarely undergo a formative (continuous improvement) or summative evaluation.

Accountability systems should be informed by well-articulated theories of action that describe the indicators, how the indicators are used to make accountability decisions, what consequences are associated with each classification, and a clear rationale for how those consequences lead to improved student performance. This session presents examples of recently completed and ongoing accountability evaluations. Importantly, this session will include the results and policy implications associated with several state evaluations.

Session Organizer:
Scott Marion, Center for Assessment

Chair:
Scott Marion, Center for Assessment

Presentations:
Calculating the Consistency of Accountability Decisions in Complex Systems
Andrew Krist

Evaluating The Accuracy and Consistency of Accountability Systems in Multiple States
Arthur Thacker, Human Resources Research Organization

How Do Intended Users Interpret and Use State Accountability Data?
Nathan Dadey, The National Center for the Improvement of Educational Assessment

Does The Accountability System Yield Results That Support State Policies?
Monica Gribben, HumRRO

Discussant:
Kyu-Ryng Hwang, New Hampshire Department of Education

111. LLM respondents in Measurement

Coordinated Paper Session
1:30 PM - 3:00 PM
Hilton Denver City Center Lower Level 2: Colorado AB

Generative AI offers an attractive alternative for developing measurement tools due to its cost-effectiveness and accessibility compared to traditional methods. Can response-level data be simulated with LLMs to possibly gain insights into human psychometrics? We will navigate through the process of respondent answer generation using Generative AI, and explore the promise of using this item-level response data to offer preliminary validity evidence, conduct item evaluation, and,

ultimately, the promise and caveats in calibrating item pools. The first paper introduces a cognitive framework/strategy for simulating respondent profiles. The approach proved effective in generating diverse and valid responses and offered initial validation results for item content creation. The second paper investigates the possibility of using fully synthetic data for item calibration within three content areas. It also discusses the influence of prompting strategies and the effectiveness of fine-tuning. The third paper further explores the power of combining synthetic and real respondent data, how Generative AI aids the incomplete data problem, and whether cost-effectiveness can be achieved by limiting the number of necessary respondents. The final paper provides a detailed case study on how to carry out the simulations and validity checks for silent sentence reading efficiency (SRE) test development purposes.

Session Organizer:
Yunting Liu, University of California, Berkeley

Chair:
Zachary Pardos, University of California, Berkeley

Presentations:
Leveraging LLM-Respondents for Item Evaluation: a Psychometric Analysis
Yunting Liu, University of California, Berkeley; Shreya Bhandari, University of California, Berkeley; Zachary Pardos, University of California, Berkeley

Developing Parallel Forms for Sentence Reading Efficiency Using LLM-based Item Response Simulator
Wanjing (Anya) Ma; Eric Zelikman, Stanford University; Jasmine E. Tran, Stanford University; Benjamin Domingue, Stanford University; Nick Haber, Stanford University; Jason Yeatman, Stanford University

Psychometric Alignment: Capturing Human Knowledge Distributions via Language Models
Joy He, Stanford University; Wanjing (Anya) Ma; Kanishk Gandhi, Stanford University; Benjamin Domingue, Stanford University; Emma Brunskill, Stanford University; Noah Goodman, Stanford University

Generative Students: Using LLM-Simulated Student Profiles to Support Question Item Evaluation
Xinyi Lu, University of Michigan; Xu Wang, University of Michigan

Discussant:
Ikkyu Choi, ETS

112. Integrating Multimodal Process Data in Digital Assessments

Coordinated Paper Session
1:30 PM - 3:00 PM
Hilton Denver City Center Lower Level 2: Denver 4

The advent of immersive digital technologies offers exciting opportunities to deepen our understanding of human cognition and develop innovative, technology-enhanced experiences for individuals. Multimodal process data, which includes detailed records of eye-tracking, conversations, gestures, body movements, neuroimaging, and more, has aroused increasing attention in digital assessments. This new type of data source provides additional layers of information for a comprehensive understanding of respondents' behaviors and decision-making process. This coordinated session highlights four novel studies in incorporating multimodal process data in digital assessments through empirical research and systematic review. The first paper presents a methodological framework that incorporates statistical and machine learning methods to explore latent constructs from eye-tracking data associated with cognitive diagnostic model. The second paper introduces functional multiple-set canonical correlation analysis on neuroimaging data to facilitate the identification of brain regions activated during working memory tasks. The third paper

highlights the use of multimodal data from eye-tracking and think aloud protocol to investigate classroom-based assessment, shedding light on the relationship between visual attention and self-regulatory behaviors. The last paper presents a systematic review on the utilization of virtual reality in immersive educational assessments and summarizes the methods to handle multimodal process data captured through immersive technologies.

Session Organizer:
Qiwei He, Georgetown University

Chair:
Qiwei He, Georgetown University

Presentations:
Exploring Latent Constructs through Multimodal Data Analysis
Shiyu Wang, University of Georgia; Shushan Wu, University of Georgia; Luyang Fang, University of Georgia; Yinghan Chen, University of Nevada, Reno

Integrating Various Sources of Functional Data with An Application to Neuroimaging Data
Heungsun Hwang, McGill University

Using Multimodal Data to Investigate Classroom Assessment
Ernesto Panadero, Dublin City University

Incorporating Virtual Reality process data in Immersive Assessments: A Systematic Review
Qiwei He, Georgetown University; Yian Hu, Viviana Luccioli, Georgetown University

Discussant:
Susan Embretson, Georgia Institute of Technology

113. Setting the Context for the Fifth Edition of Educational Measurement (Joint NCME/AERA Session)

Coordinated Paper Session
1:30 PM - 3:00 PM
Hilton Denver City Center Lower Level 2: Colorado EF

Four chapters from the forthcoming fifth edition of Educational Measurement (EM5) will be discussed in this symposium. These chapters provide important background and social context for the other chapters that appear in the volume. The chapters are Factors Influencing Education and Educational Measurement; The Sociocultural Context of Educational Assessment; Fairness in Educational Measurement: Theory and Concepts; and On the Nature of Measurement. A brief overview will be presented by the first author of each chapter, followed by a critique of the chapters provided by two EM5 reviewers. Audience participation will be invited following the discussants' comments. (Please note that this is one of two EM-5 focused sessions at the annual meeting.)

Session Organizer:
Linda Cook, ETS

Chairs:
Linda Cook, ETS
Mary Pitoniak, ETS

Presentations:
Factors Influencing Education and Educational Measurement
Linda Cook, ETS; Mary Pitoniak, ETS

The Sociocultural Context of Educational Assessment
Kadriye Ercikan, ETS; Guillermo Solano-Flores, Stanford University

Fairness in Educational Measurement: Theory and Concepts
Rebecca Zwick, University of California, Santa Barbara

On the Nature of Measurement
Derek Briggs, University of Colorado - Boulder; Andrew Maul; Joshua A. McGrane, University of Melbourne

Discussants:
Ellen Forte, edCount, LLC
Tim Moses, Buros Center for Testing

114. Graduate Student eBoard Session 4

Graduate Student Electronic Board Session
1:30 PM - 3:00 PM
Hilton Denver City Center Lower Level 2: Colorado GH

Presentations:

1. **Exploring Calculator Use Among Students with Calculation Difficulties Using Test Process Data**
   Jakob Panzer, Michigan State University; Sara Witmer, Michigan State University
   Previous research suggests many students with disabilities (SWD) do not use calculator accommodations during testing (Bouck & Bouck, 2008). This study uses test process data to explore predictors of calculator use and the relationship between calculator use and test performance among SWD who show calculation difficulties.
2. **Advanced Predictive Analytics for Equitable Education: Addressing Student Dropout and Academic Success**
   Xiaoting Zhong, University of Iowa; Stephen Dunbar, University of Iowa
   This study employs elastic-net regularized generalized linear models for feature selection and compares neural networks with logistic regression in predicting student dropout and success. With reduced computational burden, neural networks demonstrate better accuracy, highlighting machine learning's potential for wider accessibility and timely identification of at-risk students to improve educational outcomes.
3. **Investigating Assumptions of Alignment Evaluations: Is AI a viable solution?**
   Kelley Stethen, ACS Ventures; Daniel Wright, Vuk Celic, UNLV
   Confirming alignments can introduce bias in results. We investigate if knowledge of the source of the alignments influences this bias. Panelists are told whether an alignment is sourced through AI or teachers, then are asked to rate the alignment. We test if the influence of the rating varies by method.
4. **Impact of GPT-4-Based Data Augmentation on Automated Korean Essay Scoring**
   IkHyeon Hong, Inha University; Yongsang Lee, Inha University
   This study explores the impact of GPT-4-based data augmentation on the performance of automated Korean essay scoring systems. Various augmentation quantities are analyzed, comparing results from balanced and imbalanced datasets to determine the most effective approach.
5. **Automated Essay Scoring and Feedback Generation Through the Use of LLM'S**
   Doris Abroampah, University of Alberta; Tahereh Firoozi, University of Alberta; Mark Gierl, University of Alberta
   This study investigates the use of Large Language Models (LLMs), such as GPT-4, to enhance Automated Essay Scoring (AES) and feedback generation using the ASAP dataset. It explores how LLMs can replicate human scoring and provide high-quality, personalized feedback, potentially revolutionizing writing assessments by supporting learning and scalability.

6. Multi-text Semantic Feature Extraction for Personality Prediction
Xinyu Liu; Hongbo Wen, Beijing Normal university; Zhifei Li, Beijing Normal university; Yingshi Xu, Beijing Normal university
This study introduces a methodology for personality prediction via multi-text semantic feature extraction using SBERT, achieving an 89% accuracy rate in multi-class scenarios. The research compares model performances under various conditions, including text processing, individual semantic feature synthesis, data smoothing, and input features, culminating in a refined research paradigm.

7. Simulating Control Bar Impact in TOEFL Listening: AI and IRT Approaches
Jungwon Kyung, University of Massachusetts, Amherst; Craig Wells, University of Massachusetts Amherst
This study explores the impact of control bar usage in TOEFL listening assessments using IRT and machine learning models. Results from simulated test-takers reveal that the Hybrid Model (IRT + Random Forest) most accurately predicts proficiency, showing that controlled replay aids lower-proficiency learners, while excessive replay can impair test validity.

9. Addressing Class Imbalances to Optimize Rapid Guessing Classifications Through Supervised Machine Learning
Zach Baron, James Madison University; Nicolas Mireles; Yu Bao, James Madison University
Current rapid guessing detection methods prioritize response time and accuracy. This study explores how supervised machine learning models, like logistic regression and random forest, improve classification by integrating response patterns, comparing optimization metrics, and addressing class imbalances. Techniques including SMOTE and class weighting significantly improved model performance.

9. Detecting Respondents' Engagement Patterns and Differences in Sequence-based Similarity Using PIAAC Data
Xiaoxiao Liu, University of Alberta; Ying Cui, University of Alberta; Yizhu Gao, University of Georgia; Wenshuo Li, University of Alberta
This study developed a test engagement indicator based on respondents' action sequences. Using PIAAC data, three test engagement patterns (passive, balanced, high) were detected through k-means clustering. Dynamic Time Warping (DTW) revealed high sequence similarity between balanced and high groups but low similarity between passive vs. balanced and high groups.

10. Modeling Response Time: The F>C Phenomenon and the Distance-Difficulty Hypothesis
Nicole Bonge, University of Arkansas; Ronna Turner, University of Arkansas
Using 2019 TIMSS data from eighth-grade students in the United States, this study investigates two theoretical explanations for variation in response time: the F > C phenomenon and the distance-difficulty hypothesis. Results support the F > C phenomenon, the distance-difficulty hypothesis, and an interaction between the two.

11. Sequential Analysis of Keystroke Data on Coding Tasks
Kefan Yu; Min Li, University of Washington
This study aims to investigate coding action logs to identify sequential patterns among students with different experience levels. Specifically, we generated sequential actions from events and editing types, then identified behavioral patterns across varying experience levels. Preliminary results show that a student with 8 attempts had more insertions than deletions.

12. A Comparison of Latent Variable Approaches to Careless Survey Response Detection
Kari Roberts, Florida State University
Two recent approaches to detecting careless survey responses include Item Response Theory (IRT) person-fit statistics, and factor mixture models. This study will examine the performance of these two approaches in realistic conditions to help illuminate what conditions are necessary to use these two approaches to detect careless survey responses.

13. **Multi-Facets Generalizability Theory Design Using SEM for Second Language Reading Assessment**
    Junhee Park, University of Iowa
    This study demonstrated how two-facets univariate generalizability theory can be explained by the framework of structural equation modeling to illustrate the reliability and generalizability of second language reading assessment by Mplus. It showed how English leaners' score consistency changes depending on the number of items and test administrations.
14. **Exploring Generative AI Assisted Item Development in Mathematics: A Mixed-Method Approach**
    Sarah Osaro, University of Iowa; Dan Song, The University of Iowa; Catherine Welch; Stephen Dunbar, University of Iowa
    Generative artificial intelligence (AI) offers potential benefits of scale in large-scale assessment. An AI-enhanced item development approach was used to generate mathematics items aligned to state content standards, and results were evaluated against multiple criteria by subject-matter experts and reviewers using a mixed methods approach.
15. **Evaluation of IRT Chained Equipercentile Equating under the Common Item Nonequivalent Groups Design**
    Min Liang, National Board of Osteopathic Medical Examiners (NBOME); Won-Chan Lee, University of Iowa; Stuart Barnum, The National Board of Osteopathic Medical Examiners; Jinghua Liu, The National Board of Osteopathic Medical Examiners
    This study aims to evaluate the performance of the Item Response Theory (IRT) chained equipercentile equating method in comparison with three other widely used equating methods based on different investigated factors under the common-item nonequivalent groups (CINEG) design. A simulation study was developed to achieve this objective.
16. **A Methodological and Practical Exploration of Computational Propaganda Detection and Classification Models**
    Amber Hatch, University of Florida
    This thesis explores features contributing to the performance of computational models for detecting and classifying propagandistic language techniques in text, including classification levels, optimizer functions, sequence lengths, and hyperparameter variations. It also includes an application study which demonstrates the models' potential uses in educational research.

115. Sampling Design and Imputation Method Advances for Large-Scale Survey-Based Assessments

Individual Paper Session
1:30 PM - 3:00 PM
Hilton Denver City Center Lower Level 2: Denver 5-6

Chair:
Jeneve Swaby

Presentations:

On Sampling Variance Estimation of Multilevel Models under Complex Sample Design
Xiaying Zheng, American Institutes for Research; Shenghai Dai, Washington State University
This research examines sampling variance estimation in multilevel modeling (MLM) under complex sample designs. Through simulations, it compares the design-based sandwich estimator with lesser-known replication methods. Results underscore the importance of correct sandwich estimator specification, validate the replication methods as viable alternatives, and provide guidance to secondary researchers.

Handling Missing Contextual Data in Large-Scale Assessments: The Multiple Imputation Strategy
Deji Suolang, University of Michigan - Ann Arbor; Paul Bailey, American Institutes For Research; Leslie Rutkowski, Indiana University
Multiple Imputation (MI) provides an alternative and potentially enhanced method for analyzing incomplete data in large-scale assessments, compared to coding as missing that is currently being used for LSA. Findings indicate that MI delivers satisfactory quality in imputing contextual variables and effectively addresses overestimation bias in the latent assessment outcome.

Discussant:
Paul Bailey, American Institutes For Research

116. Test Security: Methods and Considerations

Individual Paper Session
1:30 PM - 3:00 PM
Hilton Denver City Center Lower Level 2: Colorado IJ

Chair:
Nick Trout, Michigan State University

Presentations:

Using Item Scores and Response Times to Detect Item Compromise in CAT
Chansoon (Danielle) Lee; Kylie Gorney, Michigan State University; Jianshen Chen, College Board
Sequential procedures have been shown to be effective methods for detecting compromised items. This study proposes and evaluates several IRT-based sequential procedures that use both item scores and response times. Results suggest the proposed single multivariate procedure is promising with improved true positive rates and reasonable false positive rates.

Item Compromise Detection Using Response Times, Scores, and Distractors
Merve Sarac, College Board
This study explores the utility of response times, in addition to item scores and distractors, in item fit analysis to enhance the detection rates of compromised items. By leveraging misfitting items, simulations showed that the false positive rate and the true positive rate improved for detecting items and examinees.

Weighted Answer Similarity Analysis
Nick Trout, Michigan State University; Kylie Gorney, Michigan State University
We develop an extended version of the ? statistic that accounts for the particular items on which similar answers are observed. The results show well-controlled Type I error rates and an increase in power when comparing the weighted ? statistic with the original unweighted version.

Evaluating Multi-Form Anchor Designs in the Presence of Item Preknowledge
Tongtong Zou, University of Wisconsin – Madison; Chansoon (Danielle) Lee; Kelly Rewley; Jerome Clauser, American Board of Internal Medicine
When administering multiple forms, having a common anchor set helps ensure each form's scores are comparable. However, having only one anchor set across all forms increases susceptibility to item exposure concerns, which can bias equating results. This study compares the performance of two anchor designs in the presence of preknowledge.

Discussant:
Scott Wood, ACT, Inc.

117. Re-humanizing Assessment: A Justice-Oriented Framework for Culturally Responsive Assessment Design (NCME Committee Organized Session)
Organized Discussion
1:30 PM - 3:00 PM
Hilton Denver City Center Lower Level 1: Penrose 1

This session will showcase a justice-oriented approach to developing culturally responsive and inclusive assessments. Experts from the Center for Measurement Justice (CMJ), WestEd, and Education First will discuss their work in creating assessments that reflect the diversity of student experiences, emphasizing equity and fairness. A panelist from CMJ will present their justice-oriented item development process, which seeks to re-humanize assessment by centering the sociocultural identities of students, particularly those from marginalized communities. Panelists from WestEd and Education First will share innovative methods used in inclusive and culturally responsive assessment design. By answering predetermined questions, the panelists will discuss strategies for overcoming challenges in this work and actionable steps to foster equitable assessment practices. This session offers practical insights for educators, researchers, and policymakers who aim to create more inclusive assessments, ensuring all students feel valued and affirmed in the assessments they take.

Chair:
Maria Hamdani, The Center for Measurement Justice

Moderator:
Thao Vo, Center for Measurement Justice

Presenters:
Maria Hamdani, The Center for Measurement Justice
Priya Kannan, WestEd
Aneesha Badrinarayan, Education First
Nisha Acharya Julien, Center for Measurement Justice

118. Supporting Social and Emotional Learning in Schools and Out-of-school-time Programs: Voices from the Field
Organized Discussion
1:30 PM - 3:00 PM
Hilton Denver City Center Lower Level 1: Penrose 2

Schools and out-of-school-time (OST) programs across the United States have adopted a variety of strategies related to social and emotional learning (SEL). Despite recent political pushback in some states and districts, SEL practices are widespread, with many educators adopting SEL in response to pandemic-related disruptions of schooling, concerns about school violence, and ongoing racial disparities in educational opportunities. Practitioners in a variety of settings have been working to develop and use SEL measures to inform their work. This session brings together leaders from two school districts and an OST network, along with researchers, in a discussion of practical strategies for SEL measurement and the factors that have supported or hindered this work. The discussion will explore specific challenges associated with ensuring that measures generate high-quality, useful data and are responsive to the experiences and needs of young people from diverse cultural contexts. The goals for this session are to raise NCME members' awareness of how schools and OST programs are measuring SEL and making sense of SEL data, to explore the potential benefits and pitfalls associated with SEL measurement, and to lay the groundwork for future research-practice partnerships.

Chair:
Laura Hamilton, American Institutes for Research

Moderator:
Laura Hamilton, American Institutes for Research

Presenters:
Brittany Greiert, Aurora Public Schools;
Sarah Killion, Aurora Public Schools
Carla Evans, Center for Assessment
John Lewis, Denver Public Schools
Daniel Read, LOR Foundation

119. New Methods and Applications of DIF
Coordinated Paper Session
3:15 PM - 4:15 PM
Hilton Denver City Center Lower Level 2: Denver 1-2

Since 1980's when DIF was proposed and formally quantified, it has always been an active research area. This coordinated session will contribute to this growing research domain by offering new methods and applications of DIF. Specifically, Paper 1 focuses on DIF in relation to treatment status. It investigates whether treatment effects are influenced by item-specific variations. This paper provides a novel perspective to evaluate and explain heterogeneous treatment effects and it illuminates how DIF may distort inferences about treatment impact. Paper 2 proposes an innovative regularization method to detect intersectional DIF. Intersectionality is a theoretical framework for studying the patterning of inequalities due to its focus on the multidimensional nature of identities. While traditional DIF focuses on two group comparisons (focal vs. reference), detecting intersectional DIF faces the challenges of a large number of groups and small sample sizes. Paper 3 examines DIF detection for items that allow multiple attempts, a feature increasingly used in online learning platforms. The paper introduces a new concept, attempt-level DIF, and compares different methods, including the likelihood-ratio test and logistic regression, to identify both overall and attempt-level DIF in multiple-attempt items. Overall, this session aims to bring together the latest advancement in DIF research.

Session Organizer:
Chun Wang, University of Washington

Chair:
Ying Cheng, University of Notre Dame

Presentations:
Testing Whether Reported Treatment Effects are Unduly Dependent on the Specific Outcome Measure Used
Peter Halpin, UNC-Chapel Hill

Detecting Intersectional DIF Using Group Pairwise Truncated Lasso Penalty
Weicong Lyu, University of Macau; Chun Wang, University of Washington; Gongjun Xu, University of Michigan

Detecting DIF for Multiple-Attempt Items
Yikai Lu, University of Notre Dame; Ying Cheng, University of Notre Dame

Discussant:
Sandip Sinharay, ETS

120. The Unintended Consequences of Practical Scoring Decisions On Score Reliability and Interpretation

Coordinated Paper Session
3:15 PM - 4:15 PM
Hilton Denver City Center Lower Level 2: Denver 3

Many decisions involved in generating scores for teachers and policymakers are technical, such as selecting an IRT model or an equating algorithm. However, many decisions are practical in nature. This panel highlights the importance of understanding how practical decisions can impact final reported scores.

The first paper examines how altering accountability system indicators can significantly affect school classifications and their reliability, even without changes in the school's actual performance. The second paper explores how scaled score reporting decisions can lead to discontinuous score distributions that fail regression discontinuity design (RDD) validity tests, which may lead researchers and the public to mistakenly believe the scores have been manipulated. The third paper applies a generalizability theory framework to show how administrative decisions influence the reliability of an SEL-focused classroom observation tool.

Collectively, these papers offer key insights for researchers and measure developers into the impact of practical decisions. They also emphasize the importance of considering both practical and technical factors when constructing scores and evaluating their reliability.

Session Organizer:
Sophie Litschwartz, MDRC

Chair:
Lily An

Presentations:
The Influence of Measurement and Sampling Error on The Precision of School Accountability Scores
Lily An

50 Ways to Fail a McCrary Test: How Valid Scaled Score Distributions Fail Common Regression Discontinuity Design Validity Tests.
Sophie Litschwartz, MDRC

Applying Generalizability Theory to the Adapted Teaching Styles Rating Scale (A-TSRS): Psychometric and Practical Considerations
Sophie Barnes, Harvard Graduate School of Education

Discussant:
Sanford Student, University of Delaware

121. New Developments in Standard Setting

Coordinated Paper Session
3:15 PM - 4:15 PM
Hilton Denver City Center Lower Level 2: Denver 4

Setting performance standards is one of the most important and consequential tasks in test development and test use, so standard setting methods must be defensible and appropriate under any context. Since new contexts emerged in the past few years, they have required advances in standard setting theory and practice. In this session, three papers showcase new methods in standard setting that are designed to satisfy the new needs or take advantage of new tools from other areas of education or technology. The first paper proposes Evidence-Based Standard Setting (EBSS) as a method that provides additional sources of positive evidence about standards decisions. EBSS

combines a series of studies to explore different cut score possibilities and how they relate to various standard setting goals. The second paper applied the Embedded standard setting (ESS) method, which is based on the principled assessment design framework, to derive cut scores and compared the ESS method to the Bookmark method. The third paper explores ways that AI can assist standard setting. Results showed that the cut scores derived from the AI-assisted approaches could offer more validity evidence than traditional methods.

Session Organizer:
Lei Wan, College Board

Chair:
Lei Wan, College Board

Presentations:
Validation Approaches for Evidence-Based Standard Setting Cut Scores
Joe Grochowalski, College Board; Amy Hendrickson, The College Board; Lei Wan, College Board; Jing Feng, College Board; Jeff Wyatt, College Board; Julie Miles, College Board

Examining Test Score Interpretations on a Computer Adaptive Assessments for Alternate Assessments
Christina Schneider; Sangdon Lim, Cambium Assessment; Gary Phillips, Cambium Assessment

A Generative AI Augmentation Approach to Standard Setting
Claudia Ventura; Joe Grochowalski, College Board; Merve Sarac, College Board; Amy Hendrickson, The College Board

Discussant:
Gregory Cizek, University of North Carolina Chapel Hill

122. Electronic Board Session 4

Electronic Board Session
3:15 PM - 4:15 PM
Hilton Denver City Center Lower Level 2: Colorado GH

Presentations:

1. **An English Proficiency Test with Spanish Directions: Helpful or Irrelevant?**
   Maura Maxfield; Jennifer Lewis, University of Massachusetts Amherst; Eduardo J Crespo Cruz; University of Massachusetts, Amherst; Stephen Sireci, University of Massachusetts Amherst; April Zenisky, University of Massachusetts, Amherst
   Can translating directions for a test of English language proficiency into learners' native language improve test-taker performance and experience? In this study we compare students' performance on and perceptions of an English proficiency test where we vary the language of the directions for answering the items.
2. **Estimating Standard Errors in Neural Network Parameterization of Item Response Theory Models**
   Denis Federiakin, Johannes Gutenberg University of Mainz; Goethe University of Frankfurt; Lidia Dobria, Wilbur Wright College; Olga Zlatkin-Troitschanskaia, Johannes Gutenberg University Mainz
   Recently, a new approach to estimating IRT models was suggested by parameterizing them as variational autoencoders with shallow decoders. While estimating person standard errors (SEs) is possible via Test Information Function, for estimating SEs of item parameters we use double backpropagation. The resulting SEs are equivalent to the traditional parameterization.
3. **Small Sample Estimation and Skill Classification Using G-DINA Model and Bayesian Regularization**
   David Arthur, University of Washington
   We propose an alternative Bayesian estimation method for the G-DINA model. Through simulation studies, we show that this method far outperforms other popular methods in terms of the accuracy

of parameter estimates, while still providing similar levels of classification accuracy of skill mastery profiles.

4. **Bayesian Region of Measurement Equivalence (ROME) Framework With Multilevel Confirmatory Factor Analysis**
   Yichi Zhang, American Institutes for Research; Hok Chio Lai, University of Southern California
   Bayesian Region of measurement equivalence (ROME) is a recently proposed framework that evaluates the practical significance of measurement noninvariance. The present study extends the Bayesian ROME framework to multilevel confirmatory factor analysis (ML-CFA) models to test MI across clusters and demonstrates the proposed procedure on the TIMSS 2007 data.

5. **Two vs. Four: Multiple-Choice Versus Binary Response Items for Detecting Misconceptions**
   Jennifer Lewis, University of Massachusetts Amherst
   A misconception is an alternative cognitive process that produces undesired reasoning for incongruent problems. Most tests measure the attainment of knowledge; however, misconception identification is vital for preventing a cascade of future systematic errors. The results from an exploration of two item types for detecting misconceptions will be presented.

6. **Exploring Attribute Hierarchies in Cognitive Diagnosis from a Graph-Theoretic Perspective**
   Hyunjoo Kim; Hans Koehn, University of Illinois at Urbana-Champaign
   This research aims to estimate attribute hierarchies from observed item responses using a graph-theoretic perspective. The key question is whether item hierarchies derived from responses can help construct the underlying attribute hierarchy. A small-scale experiment suggests that this approach may be effective.

7. **A Literature Review for Translanguaging in Classroom Math Assessment for English Learners**
   Oscar Rios, University of California – Davis; Xiaochen Xu, University of California, Davis; Margarita Jimenez-Silva, University of California - Davis; Tony Albano, University of California - Davis
   This literature review explores the growing challenge of creating valid and accessible assessments for EL in the US. In this paper we suggest that translanguaging, can be a promising approach. Translanguaging aligns with the broader frameworks of UDL, testing accommodations, and multimodal assessments, while providing specific guidance for supporting ELs.

8. **Digital course contents and assessments in classrooms: teacher usage and student outcomes**
   Jinnie Choi, Savvas Learning Company
   In blended K-8 classrooms, a digital curriculum affords teachers time-saving features such as online assignments and auto-grading of student work over the school year. Are they impactful for student success? We examine evidence in platform data to investigate which teacher behavior is associated with student learning outcomes in the course.

9. **Text-based item difficulty prediction using logistic regression models with cross-validation**
   Xuechun Zhou; Hanwook Yoo, Ascend Learning; Lucia Liu, Ascend Learning
   The time-consuming nature of field testing presents significant challenges when an item bank needs informative items for a specific ability range. Aiming to provide a guided approach to field testing, this study explores the effectiveness of using logistic regression with cross-validation for item difficulty classification based on text features.

10. **Using AI Tools to Support Scale Development**
    Kimberly Colvin, University at Albany, SUNY; Yaosheng Lou, University at Albany, SUNY; Sojung Park, University at Albany, SUNY; Gabriel Schlomer, University at Albany, SUNY; Yookyeong Park, University at Albany, SUNY
    This study is an exploration of using public AI tools such as ChatGPT, Copilot, and Gemini to support scale development. The efficacy of each tool to organize a set of items into factors that most closely align with EFA results will be compared for a variety of educational and psychological scales.

11. **Factor Retention in Ordinal Variables: A Comparison on (Robust) Fit Indices and Exploratory Graph Analysis**
    Tugay Kaçak, Trakya University; Abdullah Kılıç, Trakya University
    This study evaluated the performance of EGA (GLASSO & TMFG) and fit indices (RMSEA, TLI, CFI, SRMR, robust corrections and two cut-points) in EFA. A Monte Carlo simulation study was conducted with ordinal variables. The results indicate that robust fit indices were better, also RMSEA performed well, followed by SRMR.
12. **A nonparametric cognitive diagnosis method robust to aberrant behaviors**
    Daxun Wang, Jiangxi Normal University; Ao Jia, Jiangxi Normal University; Dongbo Tu, Jiangxi, Normal University
    This study proposes a robust general nonparametric classification (RGNPC) method for cognitive diagnosis to account for the impact of aberrant responses, with an intent of enhancing the accuracy of examinees' classifications.
13. **Detecting Gender Differential Attribute Functioning in the Subskills of a Cloze Test**
    Roghayeh Mehrazmay; Behzad Ghonsooly, Ferdowsi University of Mashhad; Jimmy de la Torre, The University of Hong Kong
    This study investigates differential attribute functioning in a Cloze test using the multiple-group G-DINA (MG-G-DINA) model to explore gender differences in mastering an attribute relative to other attributes. Results show relative mastery of two attributes (i.e., Preposition and Conjunction) in favor of males, whereas one attribute (i.e., Vocabulary) favors females.
14. **Cross-Cultural Validity of the Short Inventory of Creative Activities**
    Cayce Kantor, University of Denver; Kushmakar Baral, University of Denver; Yixiao Dong, University of Denver
    This research investigates the psychometric properties and the cross-cultural implications of the 8-item Short Inventory of Creative Activities (SICA). Rasch analysis was used to assess whether the measure was used by respondents in the intended manner and whether the scale items function differently across the German and US samples.
15. **Generating Reading Comprehension Items Using GPT and the Text Analysis Cognitive Model**
    Ayfer Sayin, Gazi University; Mark Gierl, University of Alberta
    The purpose of this study is to explore the usability of the Text Analysis Cognitive Model, which used GPT for the generation of reading comprehension items in AIG. We generate items that have similar difficulty levels using model-based AIG and GPT-4 and evaluate item quality based on field test results.

123. Novel Quantitative Methods for Classroom Assessment

Individual Paper Session
3:15 PM - 4:15 PM
Hilton Denver City Center Lower Level 1: Penrose 2

Chair:
Justice Dadzie, The University of Alabama

Presentations:

Uncovering Patterns in K-12 Argumentative Writing Through Sentence Annotations
Suhwa Han, Cambium Asessment; Amy Burkhardt, Cambium Assessment; Allison Boykin, Cambium Assessment; Frank Rijmen, Cambium Assessment; Susan Lottridge, Cambium Assessment
This study presents results of a large-scale analysis of K-12 argumentative writing through sentence annotation patterns. Using latent class analysis, the study shows the presence of eight distinct types of K-12 argumentative essays, of which information is useful for teachers to diagnose strengths and weaknesses in their students' writing.

Developing Network Psychometric Methods for Nominal Data to Analyze Student Misconceptions
Jacinta Olson; Jonathan Templin, University of Iowa
Typical educational assessment analyses only use the information of whether a student got a question correct, losing all information tied to item distractors. In this proposal, we develop a network psychometric method that uses nominal data to visually and mathematically depict potential common misconceptions within an assessment.

Using Epistemic Network Analysis to Examine Internal Structure of Learning Trajectory Measures
Melissa Gallagher, U.S. Math Recovery Council; Temple Walkowiak, North Carolina State University; Brendan Eagan, University of Wisconsin Madison; Jessica Hunt, North Carolina State University; Patrick Obot, University of Houston; Emmanuel Umeh, University of Missouri; David Shaffer, University of Wisconsin Madison
Validity evidence related to internal structure is examined using quantitative analyses. However, student mathematical understandings are theorized to progress along nonlinear learning trajectories. We present an alternative measure of students' mathematical conceptions using epistemic network analysis to examine how these conceptions connect and provide validity evidence related to internal structure.

Discussant:
Brian Gane

124. Examining Student Response Behavior in Large-Scale Survey-Based Assessments

Individual Paper Session
3:15 PM - 4:15 PM
Hilton Denver City Center Lower Level 2: Denver 5-6

Chair:
Yifan Zhang, The University of Hong Kong

Presentations:

Relationships among Effort, Motivation, and Performance on NAEP Mathematics
Blair Lehman, Brighter Research; Jesse Sparks, ETS; Jennifer Bailey, Educational Testing Service
Motivation and effort can impact performance on low-stakes assessments like the National Assessment of Educational Progress (NAEP). However, most prior research has not investigated the dynamics between test effort, domain-specific motivation, domain-general motivation, and test performance. We present findings on the dynamics between these constructs and their relationship to performance.

Persistence or disengagement? Re-examining effort through process data from TIMSS
Elena Papanastasiou, University of Nicosia; Aikaterini Gkolia, University of Nicosia; Evi Konstantinidou; University of Nicosia
This study investigates the relationship between achievement and response times for examinees who spend more time than the median on test items to examine test-taking effort. Utilizing data from the TIMSS 2019, the study introduces novel process data indicators and employs multilevel modeling for the analysis with 35 educational systems.

Drawing Blanks? How Different Item Types Impact Planned and Unplanned Item Skipping
Ella Anghel, Boston College; Matthias von Davier, Boston College
We examined whether skipping items is more likely in multiple-choice (MC), drop-down, or constructed response items. MC items were more likely to be skipped but they were skipped fewer times. Students who skipped MC items were more likely to immediately retry them, suggesting they were accidentally skipped.

Discussant:
Igor Himelfarb

125. Predicting and Reporting Outcomes in Higher Education

Individual Paper Session
3:15 PM - 4:15 PM
Hilton Denver City Center Lower Level 1: Penrose 1

Chair:
Yu-Lan Su

Presentations:

Predicting Academic Success: The Comparative Strength of HSGPA, ACT, and Demographic Factors
Edgar Sanchez
This study uses dominance analysis to assess the relative importance of high school GPA, ACT scores, and student demographics for predicting first-year college GPA (FYGPA). The study highlights the advantages of dominance analysis, including multicollinearity handling and enhanced predictor interpretability, while incorporating multiple imputation for missing data.

A Machine Learning Approach for Predicting Underrepresented Students' STEM Major Choices
Gozde Sirganci, Southern Methodist University; Yanjun Pan, Southern Methodist University; Can Mese, Indiana University
The purpose of this study is to identify the critical factors influencing underrepresented high school students' decisions to major in STEM using an exploratory machine learning approach and to take an important step in developing effective strategies to increase their participation in these fields.

Breaking Down Barriers in Medical Education Outcomes by Disaggregating Race/Ethnicity
Wendy Christensen, University of Colorado School of Medicine; Rachael Tan, University of Colorado School of Medicine; Tai Lockspeiser, University of Colorado School of Medicine
Medical education disparities by race/ethnicity are often examined using a binary categorization of whether a racial/ethnic group is under-represented in medicine. We demonstrate using three cohorts of standardized test data that disaggregating race/ethnicity information reveals important trends in outcomes that binary categorization fails to capture.

Discussant:
Bob Schwartz, National Conference of Bar Examiners

126. Using Process Data to Understand Student Response Behavior

Individual Paper Session
3:15 PM - 4:15 PM
Hilton Denver City Center Lower Level 2: Colorado AB

Chair:
Cheng Tang, The University of Georgia

Presentations:

Rapid Guessing in Continuous Medical Certification Exams: Impact and Implications
Yichong Cao, The University of Iowa; Derek Sauder, American Board of Internal Medicine
Recently, many medical certification boards have adopted continuous assessments that consist of multiple shorter assessments over a fixed time period. This study examines the impact of rapid guessing on item parameter estimation, drift and equating, examinee ability estimation, and pass/fail decisions, and also how these impacts compound over time.

Navigational Test-Taking Behaviors on a High-Stakes Admissions Assessment
Yi Yang, ETS; Guangming Ling, ETS
The purpose of the study is threefold: to categorize navigational test-taking behaviors and define various measures for navigation activities, to determine whether navigational behaviors are associated with scoring performance, and to investigate if specific navigation behaviors vary across demographic groups using a dataset of a computer-based higher education admissions exam.

Evaluating Cognitive Processes Underlying Multiple-Choice and Short-Answer Questions Using Eye-Tracking Data
Thai Ong, NBME; Brian Clauser; Su Somay, National Board of Medical Examiners; Linda Ranieri, NBME; Crystal Gorum Bing, National Board of Medical Examiners; Ulana Luciw-Dubas, National Board of Medical Examiners; Andrew Houriet, NBME; Melissa Margolis, NBME; Victoria Yaneva, National Board of Medical Examiners
We used eye-tracking technology to record eye movements of participants completing sets of items, each delivered in multiple-choice and short-answer format, and analyzed their eye movements to draw inferences about the cognitive processes used in responding to these two item formats.

Discussant:
Klint Kanopka, New York University

127. Evaluating Reliability, Precision, and Generalizability

Individual Paper Session
3:15 PM - 4:15 PM
Hilton Denver City Center Lower Level 2: Colorado CD

Chair:
Hyeri Hong, California State University, Fresno

Presentations:
Impact of Response Quality on Topic Modeling of Open-Ended Questions
Xiyu Wang; Yukiko Maeda, Purdue University; Yuxiao Zhang, Purdue University
We investigated how sample size and length of open-ended responses influence topic recovery in a Latent Dirichlet Allocation (LDA) model through simulations. By selecting realistic data conditions reflective of actual educational contexts, this study offers insights into the feasibility of using LDA in various scenarios involving open-ended response analysis.

Bootstrap Confidence Intervals for the Coefficient Omega Difference Between Two Independent Groups
Reginald Ziedzor, Amplify Education Inc; Steve Newton, Amplify Education Inc; Guoguo Zheng, Amplify Education Inc; Seyfullah Tingir
This study evaluated the performance of four bootstrap-based methods for making confidence intervals for a difference in coefficient omega between two independent groups. The results demonstrated sufficient and comparable statistical power and Type I error rates for the nonparametric bootstrap methods.

Revisiting the Importance of Distributional Assumptions for Generalizability Studies of Count Data
Adam Lekwa, Rutgers, The State University of New Jersey
Generalizability theory is frequently used in research on curriculum-based measures (CBM). Data from these measures are typically in the form of counts of correct responses. Results of this simulation study indicate that, for smaller counts (i.e., less than 20), linear and generalized linear models yield substantially different results.

Discussant:
Dongmei Li

128. Item Design and Item Difficulty

Individual Paper Session
3:15 PM - 4:15 PM
Hilton Denver City Center Lower Level 1: Mattie Silks

Chair:
Onur Ramazan, Washington State University

Presentations:

Semiotic Complexity in Computer-Administered Tests: Examining Item Design in Major Testing Programs
Eunjung Myoung, Stanford University; Guillermo Solano-Flores, Stanford University; Maria Ruiz-Primo. Stanford University; Hsiaolin Hsieh, Stanford University
Optimal design in computer-administered items allows students to focus on the content assessed rather than on interpreting and navigating the testing platform. We examined a sample of items to determine how effectively the designs of three major large-scale testing programs (NAEP, PISA, and SBAC) minimize unnecessary semiotic complexity.

Do Animated Agents Serve as Retrieval Cues in Assessments?
Christina Sondermann, German Institute for Adult Education; Hannes Schröter, FernUniversität Hagen; Martin Merkt, German Institute for Adult Education
Our online experiment investigated whether animated agents presented during learning may serve as retrieval cues in assessments. Contrary to our hypothesis, seeing the same animated agent in a learning video and the assessment did not improve information retrieval. However, visible agents reduced extraneous cognitive load and enhanced enjoyment during learning.

Using Explanatory IRT Models to Guide Item Development: Illustration with Spatial Reasoning
Daniel Oyeniran, The University of Alabama; Andrew Krist; Mehd Rajeb; Stefanie Wind, The University of Alabama; Joni Lakin, University of Alabama
Linear Logistic Test Models (LLTM) allow researchers to explore sources of item difficulty. We explored use of the LLTM to develop items with predictable difficulty. We found relatively weak replication of the LLTM estimates, reflecting the potential sample-dependence of this approach.

Discussant:
Jon Twing, HumRRO

129. Test Security: Fradulent Behavior Detection Methods

Individual Paper Session
3:15 PM - 4:15 PM
Hilton Denver City Center Lower Level 2: Colorado IJ

Chair:
Daihui Xiao

Presentations:

Automated Plagiarism Detection in Spoken Responses using Linguistic and Acoustic Features
Syed Abdul Hadi; Kyoungwon Bishop; Hacer Karamese; Sooyong Lee
This research introduces an AI-driven system to detect plagiarized responses in spoken language assessments, addressing potential use of pre-prepared or AI-influenced responses. By analyzing interpretable speech features like lexical complexity, fluency, and prosody, the system detects plagiarized responses, enhancing test security and ensuring fairer evaluations.

Comparing clustering algorithms for collusion detection using mixed-type data
Soo Ingrisone, HumRRO; James Ingrisone, Pearson VUE
Cluster analysis with mixed-type data, commonly found in testing data, is frequently challenging.

This study examines four clustering methods across 12 different simulated conditions and real data, offering insights into selecting optimal clustering strategies for identifying aberrant examinees in the mixed-type data context.

Simultaneous Detection of Cheaters and Cheating Items Using Biclustering Approach
Hyeryung Lee; Louis Roussos, Cognia; Walter Vispoel, The University of Iowa; Sandra Sweeney; Cognia
We introduce a biclustering approach to detect cheating on tests by identifying coherent subgroups of examinees and items exhibiting suspicious response patterns that is widely applicable across testing contexts. Real-data analyses demonstrate the method's superiority over other methods, with simulated data further illustrating the method's effectiveness across diverse testing situations.

Discussant:
Sergio Araneda, Caveon

130. Continuing Analyses of Frederic Lord's Randomly Parallel Testing

Coordinated Paper Session
4:30 PM - 6:00 PM
Hilton Denver City Center Lower Level 2: Denver 3

This session explores both the theoretical and practical perspectives of Lord's 1955 invention of randomly parallel testing (RPT) and its large item pool requirement. According to Lord and current research, RPT has shown substantial promise in reducing cheating, producing test scores with greater generalizability, and providing a large supply of repeatable equivalent tests. Five presentations will cover RPT from various perspectives. Principles and procedures of Generalizability Theory and Item Response Theory will be applied to the use of RPT and stratified-RPT. The papers will showcase the effects of RPT on both random and systematic measurement error. Additionally, recommended procedures will be presented to quantify psychometric properties such as dependability, reliability, and conditional standard errors of measurement. Procedures for efficiently building randomly parallel tests today, particularly those requiring large item pools, will utilize generative AI, Automatic Item Generation (AIG), and SmartItems. One paper explores a method for setting performance standards for RPT. Overall, the papers offer a broad mix of test theory, statistical procedures, data simulations, and empirical research. This session shows that improving RPT methods boosts assessment reliability and fairness, preventing cheating and creating a more equitable educational environment that ultimately benefits society.

Session Organizer:
David Foster

Chair:
David Foster

Presentations:
Applications of Generalizability Theory when Persons Take Different Random Samples of Items
Robert Brennan, The University of Iowa

Standard Setting Procedures for Automatic Item Generation-Based Randomly Parallel Testing
Tong Wu; Kristin Villanueva, UNC Charlotte; Stella Kim, University of North Carolina Charlotte; Qiao Liu

Applying Item Response Theory to SmartItems
Craig Wells, University of Massachusetts Amherst; Lisa Keller, University of Massachusetts, Amherst; Jungwon Kyung, University of Massachusetts, Amherst; David Foster

Reliability in Randomly Parallel Tests: A Practical Guide
Sergio Araneda, Caveon; Chris Foster, Caveon; David Foster; Andrew Marder, Caveon Test Security

Psychometric Properties of Scores for Individuals Taking Randomly Parallel Tests
Won-Chan Lee, University of Iowa; Stella Kim, University of North Carolina Charlotte; Seungwon Shin, University of Iowa

Discussant:
Stephen Sireci, University of Massachusetts Amherst

131. Recent Developments and Applications of Cognitive Diagnostic Models in Educational Measurement

Coordinated Paper Session
4:30 PM - 6:00 PM
Hilton Denver City Center Lower Level 2: Colorado IJ

This symposium brings together cutting-edge research on cognitive diagnostic models (CDMs), a psychometric framework aimed at providing detailed insights into students' cognitive processes. CDMs are particularly useful in educational settings, offering fine-grained feedback on student strengths and weaknesses, which can guide targeted instruction and interventions. As CDMs continue to evolve, they have expanded in application across large-scale assessments, longitudinal studies, and novel diagnostic methods. Four presentations will explore different facets of CDM advancements. The first presentation focuses on developing cognitive model-based assessment items to enhance diagnostic accuracy in large-scale assessments. The second presentation focuses on the Growth CDM, tracking skill development over time, particularly in the context of second/foreign language learning. The third presentation addresses the challenge of aligning diagnostic assessments with blueprints in a simulation study, offering solutions for ensuring content validity and accurate score interpretation. Finally, the last presentation introduces the extended MC-DINA model leveraging multiple-choice distractors to extract additional diagnostic information and providing deeper insights into student cognitive abilities. The symposium aims to foster dialogue on the future of CDM applications and their role in educational measurement. By synthesizing the results of the presentations, the symposium will provide broader implications for enhancing assessment practices and learning outcomes.

Session Organizer:
Farshad Effatpanah Hesari, Technische Universität Dortmund

Chairs:
Olga Kunina-Habenicht, Technische Universität Dortmund
Farshad Effatpanah Hesari, Technische Universitat Dortmu

Presentations:
Exploring Fine-Grained Writing Subskill Development with the Growth Diagnostic Classification Model
Hamdollah Ravand, Vali-e-Asr University of Rafsanjan; Farshad Effatpanah Hesari, Technische Universität Dortmund; Olga Kunina-Habenicht, Technische Universität Dortmund; Matthew Madison, University of Georgia

Increasing CDMs' Impact in Large-Scale Educational Assessments through Cognitive Model-Based Items
Philipp Sonnleitner, Luxembourg Centre for Educational Testing, University of Luxembourg; Lokman Akbay, Department of Educational Sciences, Istanbul University-Cerrahpasa; Turker Toker, Faculty of Educational Sciences, Department of Educational Measurement and Evaluation.Usak University; Steve Bernard, Luxembourg Centre for Educational Testing, University of Luxembourg

Diagnostic Assessment Design with Blueprint Specifications
Matthew Madison, University of Georgia; Nancy Alila, Department of Educational Psychology, University of Georgia, Athens, GA, USA

Comparing the Fit of Several CDMs to Spatial Reasoning Data
Jimmy de la Torre, The University of Hong Kong; Xiaopeng Wu, Faculty of Education, The University of Hong Kong, Hong Kong, China; Zechu Feng, The University of Hong Kong; Lin Luo, Faculty of Education, The University of Hong Kong, Hong Kong, China

Discussant:
Jonathan Templin, University of Iowa

132. Technology-based assessment of higher education students' critical online reasoning skills

Coordinated Paper Session
4:30 PM - 6:00 PM
Hilton Denver City Center Lower Level 2: Denver 1-2

This symposium presents the new technology-based assessment of students' critical online reasoning (COR) defined as multifaceted skillset at searching, evaluating, and reasoning critically from web information. The COR-assessment was developed following Evidence-Centered Design principles by Mislevy and validated according to education testing standards by AERA, NCME, APA. In this assessment, students from different disciplines responded to scenario-based tasks on topics from everyday life and the domains they studied by researching sources on the open Internet and writing mini essays. Students completed the COR-tasks from home, logging onto a virtual machine, and their web navigation and response behavior was tracked in real time. Their task processing and responses were scored using a newly developed and validated scoring scheme, including online search behavior (logdata indicators), their arguments and sources cited as evidence, and their criteria for selecting sources with regard to relevance and credibility. This session offers example insights into the open web-based performance assessment of highly complex multifaceted constructs such as COR in an authentic ecologically valid manner and presents validity evidence from studies on correlations with other variables (prior digital skills experience) and on cognitive validity (students' COR task-solving behavior from the eye-tracking and ChatGPT (sub)studies).

Session Organizer:
Dimitri Molerov, Johannes Gutenberg University Mainz

Chair:
Henry Braun, Boston College

Presentations:
Valid Assessment of Economics Students' Domain-Specific Critical Online Reasoning Skills
Marie Nagel; Olga Zlatkin-Troitschanskaia, Johannes Gutenberg University Mainz; Lukas Trierweiler, Johannes Gutenberg University Mainz; Johannes Hartig, DIPF; Philine Drake, DIPF

COR Processes – Eye-Tracking Analysis of High and Low-Performers
Andreas Maur; Ann-Kathrin Kunz, Johannes Gutenberg University Mainz; Sebastian Brückner, Anika Kohmer, Johannes Gutenberg University Mainz; Olga Zlatkin-Troitschanskaia, Johannes Gutenberg University Mainz

Students' COR task solving in economics: time-to-draft with and without ChatGPT
Dimitri Molerov, Johannes Gutenberg University Mainz; Kevin Shenavai, Johannes Gutenberg University Mainz; Denis Federiakin, Johannes Gutenberg University of Mainz; Goethe University of Frankfurt; Olga Zlatkin-Troitschanskaia, Johannes Gutenberg University Mainz; Lukas Trierweiler, Johannes Gutenberg University Mainz

Impact of vocational education on student media usage skills: a quasi-experimental design
Paul Hodes; Denis Federiakin, Johannes Gutenberg University of Mainz; Goethe University of Frankfurt; Olga Zlatkin-Troitschanskaia, Johannes Gutenberg University Mainz; Lisa Martin de los Santos Kleinz, Johannes Gutenberg University Mainz; Jennifer Gabler, Johannes Gutenberg University Mainz; DimitriMolerov ,Johannes Gutenberg University Mainz

Discussant:
Henry Braun, Boston College

133. Understanding Test-Taking (Dis)engagement: Predictors, Test Design, and Eye Movements in Low-Stakes Assessments
Coordinated Paper Session
4:30 PM - 6:00 PM
Hilton Denver City Center Lower Level 2: Denver 5-6

This session addresses the critical issue of test-taker (dis)engagement in low-stakes educational assessments, a factor essential for accurately measuring test-taker performance and enhancing the validity and practical utility of these assessments. Leading researchers from various institutions across the United States and Europe will present their findings on the dynamics of test-taking (dis) engagement. Presentation 1 investigates test takers' perceived normativity of giving effort on low-stakes assessments, showing that individual differences in norms account for variations in student behavior and performance, even after considering other known factors. Presentation 2 utilizes response-time data from the PISA 2015 Science assessment to identify key demographic, achievement, and motivational variables that predict (dis)engagement, showing that only a few were robust predictors and generalized cross-culturally. Presentation 3 examines test-taker (dis)engagement in the NAEP Mathematics assessment across different grades, showing how the interaction between test and student characteristics impact (dis)engagement. Finally, Presentation 4 utilizes an eye-tracking methodology to gain detailed insights into test-taker behavior, offering a novel approach to studying (dis)engagement beyond response-time data. The session will conclude with a discussion led by an expert discussant, followed by a moderated Q&A session to explore the implications of these findings for future research and practical applications in educational assessment.

Session Organizer:
Burcu Arslan, ETS Research Institute

Chair:
Burcu Arslan, ETS Research Institute

Presentations:
The Utility of Perceived Normative Behavior to Understand Effort on Low-Stakes Assessments
Sara Finney, James Madison University; Dena Pastor, James Madison University; Riley Herr, James Madison University

A Random-Effects Meta-Analysis of Characteristics Predicting Response-Time Effort in PISA 2015
Michalis Michaelides, University of Cyprus; Demetris Avraam, University of Copenhagen; Militsa Ivanova, University of Cyprus

Engagement across Grade Levels on a Low-Stakes, Large-Scale Assessment of Mathematics
Blair Lehman, Brighter Research; Jesse Sparks, ETS; Jennifer Bailey, Educational Testing Service

Eye Movement Data Reflect Student Performance and Potential Disengagement in Multiple-Choice Assessment
Marlit Lindner, IWM - Leibniz Institut fŸr Wissensmedien; Esther Ulitzsch, University of Oslo; Gabriel Nagy, IPN – Leibniz-Institute for Science and Mathematics Education

Discussant:
Steven Wise, EngagedMeasurement

134. Elevating Test Security in Validity Theory

Coordinated Paper Session
4:30 PM - 6:00 PM
Hilton Denver City Center Lower Level 2: Colorado AB

High-stakes testing providers expend substantial resources to deter, prevent, and detect cheating activity, which suggests that the validity of test scores for high-stakes assessments depends on the effectiveness of test security. Most test validity frameworks, however, do not explicitly include test security as a component of validity. Rather, test security is commonly ignored or assumed, a subordination that has become more apparent with the rise of at-home testing. This coordinated session aims to address this gap in research by presenting various perspectives on the role of test security in validity arguments, both for in-person and at-home tests. The discussant and presenters in this session will discuss (1) the importance of test security and remote proctoring in language assessment and how they should be integrated into validity theory; (2) how test security is handled (or not) by different models of validity; (3) how "gaming" behaviors can be classified and dealt with, both in practice and with respect to validity theory; and finally (4) how security can be incorporated into argument-based validation frameworks under an "attribution inference." Taken together, we hope this session will spark more discussion and research on the role of test security in test validity theory.

Session Organizer:
William Belzak, Duolingo

Chair:
William Belzak, Duolingo

Presentations:
Gaming Behaviors and Test Validity
Bill Bonk, Pearson

Test Security and Validation Arguments in Language Assessment
Daniel Isbell, University of Hawaii at Manoa; Benjamin Kremmel, University of Innsbruck; Jieun Kim, University of Hawaii at Manoa

Attribution: A Distinct Security-Related Inference in Argument-Based Validation
Ramsey Cardwell, Duolingo; Geoff LaFlair, Duolingo; Manqian Liao, Duolingo; William Belzak, Duolingo

Bringing Security into the Fold of Validity: A Critical Review
Micheline Chalhoub-Deville, University of North Carolina Greenboro; Mark Stevens, SAS

Discussant:
Kylie Gorney, Michigan State University

135. Using Large Language Models for Alignment and Progression Tasks

Individual Paper Session
4:30 PM - 6:00 PM
Hilton Denver City Center Lower Level 2: Colorado CD

Chair:
Suhwa Han, Cambium Asessment

Presentations:
Comparison of Pre-trained Transformer Models for Standard Alignment Study
Reese Bufferfuss, HumRRO; Hye-Jeong Choi, HumRRO; Meng Fan, UMRRO; Emily Dickinson, HumRRO

There is growing interest in large language models. However, the lack of guidance can lead to unintended consequences. This study investigates the behavior of five pre-trained transformer models in standard alignment studies. Our findings will provide the strengths and limitations of each model, offering valuable insights for their effective application.

ChatGPT is an Effective Tool for Evaluating Educational Content Alignment
Farzan Karimi Malekabadi; Pooya Razavi, Edmentum; Sonya Powers
We investigated ChatGPT's performance in detecting (mis)alignments between test items and skill statements, and the effect of model and prompt characteristics. Using a pool of 12,044 aligned and misaligned pairs, ChatGPT demonstrated high accuracy across various subjects and grades, highlighting its potential for evaluating educational content alignment at scale.

Detect Learning Progression Levels: A Novel Method Using Student Responses and Artificial Intelligence
Duy Pham, New Meridian; Trung Ngo, Northeastern University Seatle; Viet Lai, Adobe Research; Hiep Nguyen, Eastern Illinois University
We introduced a novel method that utilizes machine learning to identify students' understanding and learning levels, based on their responses to a four-part item that measures a learning progression of functions. Our accuracy rate for each part ranged from 77.6% to 92.6%. The accuracy rate for learning levels was 84.8%.

Harnessing Machine Learning to Predict Student Performance: A Longitudinal Simulation Study
Yan Zhou, Houghton Mifflin Harcourt Publishing
This study investigates the performance of machine learning and regression models in predicting student outcomes using longitudinal data. Simulations reveal that linear regression (LR) outperforms other models, while neural networks perform poorly. Results highlight the importance of data structures, emphasizing LR's reliability, especially in longitudinal educational interim assessments.

Discussant:
Youngwon Kim, Harvard University

136. CAT Item Selection

Individual Paper Session
4:30 PM - 6:00 PM
Hilton Denver City Center Lower Level 2: Denver 4

Chair:
Hyeon-Ah Kang, University of Texas Austin

Presentations:
A Comparison of Item Selection Methods for Multiple Choice Option-Based CD-CAT
Meghan Leeming, University of North Carolina Greensboro; William Stout, University of Illinois at Urbana-Champaign (Emeritus); Robert Henson, University of North Carolina Greensboro
This study proposes two item selection methods, based on correct classification rates, for Cognitive diagnostic computer adaptive testing using multinomial models. They will be compared to a multinomial extension of Shannon Entropy. Average correct classification rates for each attribute item exposure will be used to evaluate the comparison.

A Likelihood-Based Profile Shrinkage Algorithm for Cognitive Diagnostic Computerized Adaptive Testing
Xiuxiu Tang, University of Notre Dame; Ying Cheng, University of Notre Dame
Various item selection algorithms have been proposed for Cognitive Diagnostic Computerized Adaptive Testing (CD-CAT), with the goal of efficiently diagnosing examinees' strengths and weaknesses. However, these algorithms often come with significant computational costs, which can hinder their practical implementation. This study aims to simplify the item selection algorithms in CD-CAT.

A Comprehensive Comparison of Item Selection Methods in CAT for Multidimensional Forced-Choice Measures

Qin Wang; Dongbo Tu, Jiangxi Normal University; Yan Cai, Jiangxi Normal University; Junhuan Wei, Jiangxi Normal University; Jiyuan Ding

MFC-CAT integrates CAT with MFC measures, leveraging the strengths of both. This article aims to propose item selection methods based on Bayesian theory (Bayes-based methods), which is expected to improve item bank utilization rates without sacrificing trait estimation accuracy and make a comprehensive comparison among FI-based, KL-based, and Bayes-based methods.

Conditional Dependence in RT-Incorporated Item Selection Methods for CAT

Xiuxiu Tang, University of Notre Dame; Ahmed Bediwy, The University of Iowa; Justin Kern, University of Illinois at Urbana-Champaign

Despite increasing interest in using response time (RT) in item selection via the hierarchical model (van der Linden, 2007), the assumption of conditional independence between RT and response accuracy (RA) may be violated. This study investigates the robustness of various RT-incorporated item selection methods for CAT under conditional dependence.

Discussant:
Kevin Cappaert, Curriculum Associates

137. Content, Item, and Test Development

Individual Paper Session
4:30 PM - 6:00 PM
Hilton Denver City Center Lower Level 1: Penrose 2

Chair:
Jaylene Patterson, University of Kentucky

Presentations:

What Do Test-takers Want? Examining the Difficulty of CAT

Paulius Satkus; Kyung (Chris) Han, GMAC

We surveyed test-takers' perceptions of computerized adaptive test (CAT) difficulty and evaluated new approaches for controlling test difficulty while preserving psychometric integrity. Survey results indicate that test-takers aim for higher success rates than traditional CATs. Our simulations suggest that a balance can be achieved by lowering test difficulty to 60-70%.

Using Content Experts' Comparative Judgment to Estimate Item Parameters

Lanrong Li, MetaMetrics, Inc.; Matthew Gushta, MetaMetrics, Inc.

This study evaluated reading item difficulty estimates resulting from expert comparative judgments against those based on student data. Item difficulty recovery by the comparative judgment method is best when the content is well aligned and the construct of measurement is narrowly defined (e.g., reading comprehension).

Measuring Word Reading Using a Linguistically Fair Spanish Lexical Decision Task

Julian Siebert, Stanford University; Mia Fuentes-Jimenez; Wanjing (Anya) Ma; Ana Saavedra, Graduate School of Education, Stanford University; Carrie Townley-Flores, Graduate School of Education, Stanford University; Jason Yeatman, Stanford University

We discuss the development of the ROAR-Palabra, a novel self-administered Spanish lexical decision task for use in linguistically diverse populations. The task has high reliability, and scores are highly predictive of conventionally assessed word reading. We assess parameter invariance and differential item functioning analysis using data from Colombia and California.

¿Sabes Que? A deep dive into Spanish item design in reading screening

Lillian Duran, University of Oregon; Mónica Zegers, CSF; Cengiz Zopluoglu, University of Oregon; Francesca Pei, University of California, San Francisco; Marilu Gorno Tempini, University of California, San Francisco

This presentation will focus on critical considerations in Spanish item development in reading screening measures including nonword reading and spelling. Item development blueprints will be shared that were based in Spanish syntax, phonology, and orthography as well as item level statistics and coverage of the ability scale.

Discussant:
Benjamin Williamson, CYBER.ORG

138. Psychometric Overlord Auditions: Take 1

Organized Discussion
4:30 PM - 6:00 PM
Hilton Denver City Center Lower Level 2: Colorado GH

One of the most critical aspects of any successful NCME conference is the exchange and debate of ideas among NCME members. This invited poster session is an opportunity to engage in discussion, debate, and evaluate innovative ideas from leaders across the measurement community. The speakers for this session includes NCME leaders (past presidents and members of the NCME Board), leading educators, and veteran practitioners in the educational measurement and the professional credentialing communities. All speakers will address the same question:
When you are appointed Psychometric Overlord for the day, what will be your first three activities?

Join us for this session and get ready to discuss the innovative, intriguing, and possibly off-the-wall ideas. Session attendees will have the opportunity to vote for their favorite ideas (and perhaps those not so favorite) and glamorous and flashy prizes will be awarded to the most popular ideas.

Chair:
Andrew Wiley, ACS Ventures, LLC

Moderator:
Susan Davis-Becker, ACS Ventures, LLC

Presenters:
Chad Buckendahl, ACS Ventures, LLC
Jennifer Dunn, Pearson
Ellen Forte, edCount, LLC
Joshua Goodman, National Commission on Certification of Physician Assistants
Kristen Huff, Curriculum Associates
Neal Kingston, University of Kansas
Susan Lottridge, Cambium Assessment
Hillary Michaels, HumRRO
Michael Rodriguez, University of Minnesota
Amin Saiar, Board of Pharmacy Specialties
Ye Tong, National Board of Medical Examiners
Michael Walker, HumRRO
Zachary Warner, New York State Education Department
Ada Woo, Ascend Learning

139. Scaling, Linking, and Equating with Test Redesign (SIGIMIE-Organized Session)

Organized Discussion
4:30 PM - 6:00 PM
Hilton Denver City Center Lower Level 1: Penrose 1

With the acceleration of knowledge updating and the advancing technology, testing programs are more frequently expanding the capability of their assessment to improve the measurement of the intended construct and/or to provide better testing experience for target population. Example changes in test redesign include updating contents, redefining specification, evaluating higher cognitive

abilities/skills using innovative item types, the self-growing item pool from automatic item generation, and the shorter administration time in adaptive testing. To ensure the success of a test redesign, it is challenging but critical to keep the score use interchangeably and/or interpretably. This session brings together practitioners from four recently redesigned large-scale standardized testing programs to share their practical considerations, issues, and solutions in scaling, linking, and equating during the test redesign process. Open-floor discussion is planned for exchanging insights and guidance benefiting current and future redesigned assessments and researchers of interests.

Chair:
Min Wang, College Board

Moderator:
Jinghua Liu, The National Board of Osteopathic Medical Examiners

Presenters:
Jaime Malatesta, GMAC
Dongmei Li
Joe Betts, NCSBN
Weiwei Cui

140. From Principles to Practice: Operationalizing Responsible AI in Educational Measurement
Organized Discussion
4:30 PM - 6:00 PM
Hilton Denver City Center Lower Level 2: Colorado EF

While high-level guidelines on the responsible use of AI provide foundational principles, the true challenge lies in the implementation. This panel discussion aims to bridge the gap between theory and practice by offering concrete, actionable insights for the ethical and effective application of AI in educational measurement. Focusing on the “how-to” of responsible AI, our discussion will cover practical strategies for integrating AI within educational measurement applications, techniques for identifying and mitigating bias in AI algorithms, and methods for ensuring data privacy and security. Additionally, panelists will discuss maintaining transparency and accountability in AI applications and exploring approaches to ensure AI is used to improve the equitable educational impact and integrity of educational measurement systems.

Designed for experts in educational measurement with varying levels of familiarity with AI, this panel aims to provide an accessible yet comprehensive roadmap for responsible AI implementation. Attendees will have the opportunity to share their thoughts, ask questions, and engage directly with panelists. The session will focus on providing practical knowledge to move beyond high-level concepts, to support the effective application of AI for improving the validity of assessments in which AI technologies are employed.

Chair:
Matthew Johnson, ETS Research Institute

Moderator:
Matthew Johnson, ETS Research Institute

Presenters:
Matthew Johnson, ETS Research Institute
Kadriye Ercikan, ETS
Odis Johnson, Johns Hopkins University; ETS Research Institute
Danielle McNamara, Arizona State University

141. States' Are Trying to Make Assessments More Instructionally Useful: How's It Working?

Coordinated Paper Session
8:00 AM - 9:30 AM
Hilton Denver City Center Lower Level 2: Denver 4

State leaders have heard complaints from local educators about state tests' lack of instructional usefulness. Simply being truthful about the limitations of statewide assessments has not quelled the concerns.

Therefore, states have tried to meet these needs in a few different ways. Starting in about 2014-2015, states started procuring their own interim or benchmark assessments along with the summative assessment. The rationale is that by providing interim and other closer-to-the-classroom assessments, state leaders can support assessments aligned with state standards and hold students to similar expectations as they are on the summative assessment.

More recently, several states have pursued through-year test designs (e.g., Dadey et al., 2023) in an effort to provide instructionally relevant information to teachers and students throughout the school year (among other purposes). However, the likelihood of educators effectively using the through-year results is contingent upon the test's design and, subsequently, how educators make sense of the through-year score reports.

This session brings together leaders in conceptualizing and evaluating instructional usefulness and researchers studying the utility of state-sponsored benchmark assessments and through-year score reports, respectively, to examine the opportunities and challenges faced by state leaders when trying to improve the instructional utility of their assessment systems.

Session Organizer:
Scott Marion, Center for Assessment

Chair:
Scott Marion, Center for Assessment

Presentations:
Evaluating the Instructional Usefulness of Assessments
Carla Evans, Center for Assessment

Investigating Teacher Use of Benchmark Modules to Improve Instruction in Utah
Kyla McClure, University of Colorado Boulder

Understanding the Potential Instructional Uses of Through-Year Assessments in Montana
Mi Jin Park, Graduate Center, City University of New York

Teachers Are Not Dumb! Assessment is Hard. Lessons from Behavioral Economics
Scott Marion, Center for Assessment

Discussant:
Nathan Dadey, The National Center for the Improvement of Educational Assessment

142. Recent Theoretical AI Developments in Educational Assessment

Coordinated Paper Session
8:00 AM - 9:30 AM
Hilton Denver City Center Lower Level 2: Denver 5-6

This session summarizes recent advancements in Artificial Intelligence (AI) for educational assessment, focusing on four key studies. The first study discusses the use of Generative Language Models (GLMs) combined with Retrieval Augmented Generation (RAG) for scoring short-answer questions, demonstrating improved accuracy through a novel scoring pipeline. The second study evaluates the efficacy of ChatGPT-4o in scoring essays and short-form constructed responses, revealing that while it shows promise, it falls short of human accuracy in some cases, particularly for high-stakes assessments. The third study explores the integration of handcrafted NLP features with deep-learning models like RoBERTa for Automated Essay Scoring (AES), proposing a two-step process that enhances prediction accuracy and model explainability. The final study examines the diagnostic accuracy of the MI Write Automated Writing Evaluation (AWE) system in predicting English Language Arts (ELA) proficiency among middle-school students, highlighting the potential and challenges of AI-based screening for diverse student populations. Collectively, these studies illustrate the growing potential of AI in educational measurement while also emphasizing the need for ongoing refinement to address issues of accuracy, bias, and fairness across various applications.

Session Organizer:
Mark Shermis, Performance Assessment Analytics, LLC

Chair:
Jill Burstein, Duolingo

Presentations:
Generative Language Models with Retrieval Augmented Generation for Automated Short Answer Scoring
Zifan Wang, Syracuse University; Christopher Ormerod

Using ChatGPT-4o to Score Essays and Short-Form Constructed Responses
Mark Shermis, Performance Assessment Analytics, LLC

Incorporating NLP features into Transformer-Based AES Models
YoungKoung Kim, The College Board; Tim Moses, Buros Center for Testing

Predicting ELA Proficiency for Middle-grade ELL and non-ELLs with Automated Writing Evaluation
Fan Zhang, University of Delaware; Joshua Wilson, University of Delaware

Discussant:
Corey Palermo, Measurement Incorporated

143. Innovations in Latent Space Item Response Modeling: Diagnosing Hidden Dependencies for Fairness

Coordinated Paper Session
8:00 AM - 9:30 AM
Hilton Denver City Center Lower Level 2: Colorado CD

Latent space models are gaining increasing attention in item response theory due to their flexibility and visualization advantages. The latent space item response model (LSIRM; Jeon et al., 2021) offers a novel IRT framework that captures hidden conditional dependencies and heterogeneity among test items and respondents. By introducing a novel latent space term into traditional item response theory models, LSIRM projects respondents and items onto a low-dimensional interaction map, where the distances between respondents and items represent unobserved interactions that might be neglected by standard IRT models while accounting for item features and person traits. This session

brings together five studies extending and applying LSIRM to address critical challenges in modern educational measurement. The papers explore LSIRM's ability to provide insights into response accuracy and response times, cheating behaviors, Likert scale data, and method effects, advancing both theory and practice for fairer, more accurate assessments. Through empirical studies and simulations, the session demonstrates how LSIRM captures nuanced respondent-item interactions, uncovers hidden conditional dependence patterns, and enhances diagnostic information, contributing to fairer educational assessments. These studies emphasize LSIRM's capacity to support socially impactful measurement practices, improving the accuracy and fairness of tests and assessments and advancing educational measurement to benefit society.

Session Organizer:
Jinwen Luo, UCLA

Chair:
Minjeong Jeon, UCLA

Presentations:
Conditional Dependence across Slow and Fast Item Responses: A LSIRM Approach
Nana Kim, University of Minnesota

Capturing Item Preknowledge with a Response Time Latent Space Model Framework
Yingshi Huang, University of California - Los Angeles

Interaction Maps for Diagnosis: How Can LSIRM Inform Us about the Respondents?
Jinwen Luo, UCLA

Ordinal Latent Space Item Response Models for Likert Scale Assessments
De Carolis Ludovica, University of Milano-Bicocca

Investigating Method Effects of Wording Valence: A Latent Space IRT Approach
Biao Zeng, Beijing Normal University

Discussant:
Matthew Johnson, ETS Research Institute

144. Fostering Equity In College Admissions: Test Reforms In Chile

Coordinated Paper Session
8:00 AM - 9:30 AM
Hilton Denver City Center Lower Level 1: Penrose 1

The present session focuses on efforts that are being taken to increase the diversity of the student population that takes college admissions tests and applies to colleges within Chile's centralized higher education admission system. In response to concerns about the pertinence, usefulness, and fairness of admissions tests, efforts have been underway for the past years to shift from broad content-based admissions tests to tests that emphasize competency with more targeted content.

The first presentation will examine the diversification of the population participating in the centralized admission system. The second will focus on a methodological change in assessing the relationship between test scores and academic performance. The third presentation will explore the shift from norm-referenced to criterion-referenced approach to score reporting using a construct mapping approach (Wilson, 2023). The fourth presentation will explore how future users interpret and anticipate uses of performance levels.

This proposal aims to contribute to the discussion on the challenges faced by university admis-sions systems that rely on standardized tests, as well as contribute to the development of best practices that can help us overcome those challenges, emphasizing the potential of diagnostic data to promote equity and inclusion in higher education.

Session Organizer:
Daniela Jimenez, DEMRE - Universidad de Chile

Chair:
Daniela Jimenez, DEMRE - Universidad de Chile

Presentations:
The Quantitative Journey Through Chilean University Admission System
María Inés Godoy, DEMRE - Universidad de Chile; Varas María Leonor, DEMRE - Universidad de Chile; Eduardo Alarcón, DEMRE - Universidad de Chile; Carlos Cayumán, DEMRE - Universidad de Chile

Characterizing Heterogeneity In Predictive Relationships Using A Beta Mixture Regression Approach
Eduardo Alarcón, DEMRE - Universidad de Chile; Dany López, DEMRE - Universidad de Chile; María Inés Godoy, DEMRE - Universidad de Chile; David Torres Irribarra, Pontificia Universidad Catolica de Chile

Developing Performance Levels For PAES: The Case Of The Maths Test
Daniela Jimenez, DEMRE - Universidad de Chile; Sandra Cruz, DEMRE - Universidad de Chile; Emilia Seissus, Institut d'études Politiques de Paris; David Torres Irribarra, Pontificia Universidad Catolica de Chile; Juan Cortes, DEMRE - Universidad de Chile; María Cristina Morales, DEMRE - Universidad de Chile

Performance Levels In Context: How Do Users Interpret And Project Their Use?
Bruna Sciolla, DEMRE - Universidad de Chile; Cristian Saavedra, DEMRE - Universidad de Chile; Yulan Sun, DEMRE - Universidad de Chile; Catalina Henríquez, University of Colorado – Boulder; Maximiliano Herrada, DEMRE - Universidad de Chile

Discussant:
Derek Briggs, University of Colorado - Boulder

145. Investigating Validity of Learning Progressions in Computational Thinking (CoT)

Coordinated Paper Session
8:00 AM - 9:30 AM
Hilton Denver City Center Lower Level 2: Colorado AB

Over the past several years, the National Math and Science Initiative (NSMI) and UC Berkeley researchers have collaborated to develop assessments in algebraic thinking, statistical thinking, and computational thinking for high school and middle school students. This coordinated set of papers reports on validity studies conducted within the project, focusing on our latest domain, computational thinking (CoT). The CoT assessment has been developed using the four-building block approach and the BEAR Assessment System (BAS, Wilson 2005). Our assessment framework comprised primarily of three dimensions: a) design, b) implement, and c) analyze/evaluate a computational solution, with a recent addition of a 4th dimension: d) conceptualize an applied problem. The first paper explores pre-post change patterns across the three primary CoT dimensions to investigate the validity of our hypothesis about learning progressions (LPs). The second paper provides evidence of measurement invariance among important student subgroups such as gender and ethnicity. The third paper investigates whether the hypothesized learning progression in each CoT dimension holds for the student subgroups. The last paper reports on the degree to which we succeeded at developing selected-response items that are of similar difficulty as their constructed-response counterparts.

Session Organizer:
Yukie Toyama

Chair:
Karen Draney

Presentations:

Investigating Pre-post Change Patterns in a Multidimensional Measurement of Computational Thinking (CoT)
Yunting Liu, University of California, Berkeley; Yukie Toyama; Mark Wilson, University of California, Berkeley

Examining the Measurement Invariance in the Computational Thinking Assessment
Richard Brown, West Coast Analytics; Emily Brown, University of North Texas

Examining the Stability of Learning Progression(s) in the CoT Assessment
Richard Brown, West Coast Analytics; Emily Brown, University of North Texas

Comparing the Difficulty of Selected-Response and Constructed-Response Items
Xingyao Xiao, University of California, Berkeley; Perman Gochyyev, University of California, Berkeley Mark Wilson, University of California, Berkeley

Discussant:
Frank Rijmen, Cambium Assessment

146. Electronic Board Session 5

Electronic Board Session
8:00 AM - 9:30 AM
Hilton Denver City Center Lower Level 2: Colorado GH

Presentations:

1. **Automated Difficulty Prediction for Binary and Multi-Scored Items**
   Ye Lin, Ascend Learning, Xiuhan Chen, University of Missouri; Kari Hodge, Ascend Learning
   The study employed machine learning to forecast the difficulty levels of both binary and multi-scored items. It examined linguistic and cognitive features as well as word embeddings from item stems and answer choices, evaluating various algorithms. The findings have the potential to enhance test development and pretesting processes.
2. **Comparison of Linking Items Displacement Criteria in Vertical Scale Development**
   Hongwook Suh, Cambium Assessment; Elizabeth Ayers-Wright, Cambium Assessment; Mario Gonzalez, Cambium Assessment
   Selection of linking items based on different criteria results in impacts to developing scaling constants and representativeness of vertical scales reflecting genuine grade-to-grade growths and variabilities among grades. Different vertical linking item selection methods and designs will be compared by analyzing simulated data generated from Grade 3-8 mathematics test data.
3. **Preliminary Development of a Student-Completed Behavior Rating Scale Measuring Future Skills**
   Jennifer Robitaille, Riverside Insights; Emily Parker; Sara Prewett; Evelyn Johnson, Riverside Insights
   This study describes the preliminary development of a student-completed measure of key future skills, such as adaptability and critical thinking, leveraging a widely used assessment of social-emotional competence and efficient real-time data collection on new items. This tool will help prepare students to succeed in school and beyond graduation.
4. **The Effect of Rating Scale Length on Dimensional Structure: A Network Approach**
   Chia-Lin Tsai, University of Northern Colorado; Samantha Estrada, The University of Texas at Tyler; Julia Fulmore, Fulmore; Nai-En Tang, National Board of Chiropractic Examiners
   This study aims to evaluate the impact of rating scale length on the dimensional structure of self-report measures using an experimental design and a network approach. Our preliminary findings showed that the rating scale lengths do not impact the dimensional structure of a self-report measure.

5. Using Minimum Average Partial Test to Analyze Residuals and Examine Unidimensionality
Tianshu Pan
Velicer (1976) proposed Minimum average partial (MAP) test to determine dimensionality in factor analysis. To avoid these limitations of applying MAP to the phi or tetrachoric correlations, we explore the feasibility of MAP applying to the raw and standardized residuals respectively to examine unidimensionality by a simulation study. Velicer (1976) proposed.

6. Language and Academic Growth Trajectories of English Learners
Yen Vo
This study examines the growth trajectories of English learners in English language proficiency, ELA, and math from Grades 6 to 11. Results show that growth patterns vary by subject area and demographic characteristics. Educators and policymakers should consider these dynamics when designing instructional strategies and support systems for English learners.

7. Comparing Data-Based Decision Making (DDM) Competency in US and Korean Students
Hyemin Park, UC Berkeley; Jihye Shin, Guwol Girl's Middle School
This study examines differences in Data-Based Decision Making (DDM) competency between US and Korean middle school students using pre- and post-test results. By analyzing progression across six waypoints (DDM0 to DDM5), we explore the influence of cultural and educational systems on students' competency.

8. Using Multilevel Modeling to Assess Differential Predictive Validity in Graduate Admissions Tests
Chris Patterson; Jonathan Templin, University of Iowa
Many graduate admissions tests claim to predict graduate school success. To interrogate this claim, this study explores differential predictive validity of graduate school admissions test scores. Done with multilevel modeling, this paper will also discuss and show the application of a seldom-used method in differential predictive validity research.

9. Student-Generated Questions and Goal Orientation in Literacy Assessment
Emily Shin, University of Toronto; Angelie Ignacio, University of Toronto; Eunice Jang, University of Toronto
This study explores how students' reading abilities and goal orientations predict their questioning skills. Questioning ability was predicted by both reading ability and goal orientation. Our findings offer key insights into literacy development, aiming to improve literacy acquisition and assessment.

10. The Impact of Weighting Methods on Norm Score Accuracy
Tong Wu; Stella Kim, University of North Carolina Charlotte; Hyeonjoo Ho, Riverside Insights; JongPil Kim, Riverside Insights
The purpose of the study is to evaluate the impact of six weighting methods on the accuracy of norming. The study investigates the relative performance of these weighting methods in conjunction with three norming techniques under various simulation conditions.

11. Impact of Pretest Item Administration and Calibration Method on Parameter-Estimates in MST
Hong Chen, University of Iowa; Nikole Gregg, National Commission on Certification of Physician Assistants; Christiana Akande, National Commission on Certification of Physician Assistants; Yanlin Jiang, NCCPA; Andrew Dallas
This proposal explores pretest item calibration in multistage testing, evaluating the robustness of scaling methods on ICC recovery when using varying proportions of common pretest items across adaptive modules. Simulations reveal that SG-SC consistently performs best, with MG-FIP outperforming SG-FIP. LRG's performance is significantly influenced by pretest item administration approaches.

12. Socioeconomic achievement gap in writing and assessment moderators: A meta-analysis
Carolina Castillo, Pontificia Universidad Católica de Chile
This study analyzes the socioeconomic achievement gap in students' writing abilities across various educational levels, using a meta-analysis of 139 effect sizes. Results indicate the gap is moderated by factors such as data source, rubric type, writing genre, and student characteristics. The findings emphasize fairness in educational writing assessments.

13. Validation Practices for Dynamic Measurement
Yixiao Dong, University of Denver; Denis Dumas, University of Georgia
Dynamic Measurement (DM) tools have been used for capturing developing constructs, such as learning potential and cognitive malleability. After a comprehensive review of the DM practices, we identify practical challenges in validating DM measures, and then propose an expanded validation framework to offer guidance for validating these measures.

147. Using Large Language Models within Item Development Workflows

Individual Paper Session
8:00 AM - 9:30 AM
Hilton Denver City Center Lower Level 2: Colorado IJ

Chair:
Henry Makinde

Presentations:

AI-generated Classification Rubrics in the Process of Item Validation
Yoav Bergner, New York University; Yu Wang, New York University; Madhumitha Gopalakrishnan, New York University
The potential of generative language models to evaluate assessment items is still uncertain, particularly when it comes to complex constructs like collaborative problem-solving. We explore whether a multi-agent workflow in which agents produce item evaluation rubrics before applying them increases the agreement with human raters and produces sound reasoning.

Exploring the Use of Generated Items in Test Equating
Ayfer Sayin, Gazi University; Mark Gierl, University of Alberta
The study aims to explore the use AIG in test equating within nonequivalent groups design. By comparing generated-items with traditional anchor-items across various equating methods, the results suggest that generated items are a promising alternative, particularly in contexts where using the same items is challenging for security or other constraints.

Finding Words Associated with DIF: Predicting and Describing DIF using LLMs
Hotaka Maeda, Smarter Balanced; Yikai Lu, University of Notre Dame
We built a large language model that predicts differential item functioning (DIF) against female or male examinees from the item text, and identify specific words that were associated with DIF. The model can help us understand how DIF manifests, and potentially provide unique feedback for improving item development.

Leveraging Large Language Models to Predict Item Difficulty
Mingfeng Xue, University of California, Berkeley; Suhwa Han, Cambium Asessment; Allison Boykin, Cambium Assessment; Frank Rijmen, Cambium Assessment
We evaluate the efficacy of three large language models (LLMs) — RoBERTa, DeBERTa, and ELECTRA for predicting item difficulty in different settings. Fine-tuning improves accuracy over pre-trained models, with DeBERTa performing best. Incorporating linguistic features and item attributes offers slight enhancements. The research underscores LLMs' potential to advance educational measurement.

Discussant:
Guher Gorgun, University of Alberta

148. Enhancing and Extending CAT

Individual Paper Session
8:00 AM - 9:30 AM
Hilton Denver City Center Lower Level 2: Denver 3

Chair:
Montserrat Beatriz Valdivia Medinaceli

Presentations:

Enhancing Bayesian Online Calibration in Computerized Adaptive Testing with Collateral Information
Ziyu Fan, The University of Texas at Austin; Seung Choi, The University of Texas at Austin
This study presents an enhancement to online item calibration within Computerized Adaptive Testing by incorporating collateral information from related assessments. Using a Bayesian framework, we improve ability estimation, comparing random and D-optimal methods for field-test item assignment. Our findings reveal significant improvements in calibration accuracy, particularly in small sample scenarios.

Optimizing Test Completion in Strictly-Timed CATs: A Post-Hoc Simulation
Nour Armoush, University of Alberta; Joyce Xinle Liu, University of Alberta; Okan Bulut, University of Alberta
In this post-hoc simulation study, we incorporated response times into item selection in computerized adaptive testing (CAT). Linear optimization was applied for test-takers at risk of not completing the test, resulting in higher completion rates and improved ability estimates, particularly for low-ability test-takers. We offer practical insights for strictly-timed CATs.

Adaptive Mastery Detection in Intelligent Tutoring Systems: a Two-Stage Approach
Sangbeak Ye, Florida Atlantic University; Jimmy de la Torre, The University of Hong Kong
Recent psychometric advances highlight the need to integrate learning behavior into ITS assessments. This study introduces a two-stage procedure to enhance mastery detection by adaptively administering items. Stratifying item banks by item properties, the method accelerates learning and reduces detection delays at a fixed false detection rate.

Discussant:
Laurie Davis, Curriculum Associates

149. Modern Types of DIF

Individual Paper Session
8:00 AM - 9:30 AM
Hilton Denver City Center Lower Level 2: Denver 1-2

Chair:
Haeju Lee

Presentations:

Modeling Variational DIF for Different Ability Levels with Moderated Mediation Approach
Gozde Sirganci, Southern Methodist University; Yusuf Kara, University of Miami ; Michael Fong, Southern Methodist University; Akihito Kamata
This study proposes a differential item functioning detection technique that can examine the existence and magnitude of DIF for varying levels of ability by extending the procedure suggested by Montoya and Jeon (2020). Variational DIF is tested by combining multiple indicators multiple causes, moderated mediation, and the Johnson-Neyman procedure.

A Comparison of Methods to Evaluate Complex Sources of DIF
Haeju Lee; Sijia Huang, Indiana University Bloomington; Dubravka Svetina Valdivia, Indiana University
In this study, we review and compare several methods for evaluating complex forms of DIF, including logistic regression (LR)-based, IRT-based DIF detection methods, as well as LR-based and IRT-based regularization methods. The performance of these methods is compared through a simulation study and empirical data analysis.

Beg to DIFfer: Examining Multiple Testing for Intersectionality
Lily An; Edward Kim, Bentley University
Recent attempts to involve intersectionality in DIF analyses consistently overlook interdependence patterns. Our simulations show that canonical p-value correction methods are inappropriate whenever identities are unevenly distributed across multiple dimensions; more so when DIF stems from intersectional, rather than sums of constituent, identities. We provide recommendations and remedies.

Addressing Intersectional DIF through Moderated Nonlinear Factor Analysis
Joseph Kush, James Madison University
While traditional multigroup approaches to measurement invariance examine discrete groups in succession, moderated nonlinear factor analysis (MNLFA) considers the moderating effects of multiple covariates simultaneously. Using simulated item-response data, results indicated that even with non-ignorable DIF, an appropriately specified MNLFA model accounting for intersectional DIF produced unbiased factor scores.

Discussant:
Tony Albano, University of California - Davis

150. Socioculturally Responsive Assessment and Measurement

Individual Paper Session
8:00 AM - 9:30 AM
Hilton Denver City Center Lower Level 1: Penrose 2

Chair:
Alexis Oakley

Presentations:
Validating an Academic Acculturation Scale for Youth
Kyndra Middleton, Howard University; Wallace Pinto; De'Ani Costello, Howard University
The current study validates an adapted academic acculturation scale for middle and high school students and provides students with culturally relevant and "regular" items to gain a deeper understanding of whether a relationship exists between academic acculturation and performance on culturally (ir)relevant items.

Developing a Measure of Teaching for Equity Using the Culturally Responsive Teacher Evaluation Framework

Megan Welsh; Kayce Mastrup, California State University, Monterey Bay; Liam Aiello, University of California, Berkeley; Kristen Pilner Blair, Stanford University; Sergio Sanchez, University of California, Davis; Ciara Thomas Murphy, University of California, Davis

Frameworks for developing culturally responsive assessments of teaching for equity, social justice, and diversity have been developed, but not applied. This paper shares efforts to apply Chang and Cochran-Smith's (2022) culturally responsive teacher evaluation framework to the development of a simulation-based assessmnet of using discussion-based teaching methods in diverse secondary English classsrooms.

Evaluation of a National-level Mentorship Program for Educational Measurement Doctoral Students

Roti Chakraborty, Georgia State University, American Institutes for Research; Joseph Rios, Center for Measurement Justice; Jennifer Randall, University of Michigan; Jennifer Rodriguez, University of Michigan

This mixed-method study aimed to evaluate a national-level mentorship program for Educational Measurement minoritized doctoral students in the U.S. Findings suggest that students experienced effective mentor-mentee relationships and various satisfactory support. In response to students' expectations, the program will implement new initiatives like instructional workshops and online learning communities.

Discussant:
Magdalen Beiting-Parrish, Federation of American Scientists

151. Testing Standards Revision: Focus on Applications of Measurement (Joint NCME/AERA Session)

Organized Discussion
8:00 AM - 9:30 AM
Hilton Denver City Center Lower Level 2: Colorado EF

This is the second of two joint sessions that will update the members of AERA and NCME and provide an opportunity for input on the Standards. Specifically, this session will address the Applications section of the Standards, with a focus on test score interpretation, use, and misuse, as well as the role of unintended consequences. The Joint Committee (JC) received its charge on August 21, 2024, has been meeting with experts, and is working on the Standards revision. In this second session, participants will receive an update on the process and hear from members of the JC regarding their work and any current challenges being addressed by the committee in the area of applications. The majority of the session is intended to hear from participants about their thoughts on the ongoing revision, in reaction to the updates or regarding their own questions, hopes, or concerns for the revision.

Chair:
Michael Rodriguez, University of Minnesota

Moderator:
Kristen Huff, Curriculum Associates

Presenters:
Michael Rodriguez, University of Minnesota
Kristen Huff, Curriculum Associates
Ye Tong, National Board of Medical Examiners
Andy de los Reyes
Chad Buckendahl, ACS Ventures, LLC
Ellen Forte, edCount, LLC

Laura Hamilton, American Institutes for Research
Qiwei He, Georgetown University
Nathan Kuncel
Cara Laitusis, Center for Assessment
Maria Marquine
Rochelle Michel
Michael Russell, Boston College
Pohai Kukea Shultz, University of Hawaii - Manoa
Stephen Sireci, University of Massachusetts Amherst
Stephen Stark, University of South Florida
Mark Wilson, University of California, Berkeley
Frank Worrell

152. Leveraging Large-Language Models for Operational Large-Scale Assessments: Opportunities and Challenges

Coordinated Paper Session
9:45 AM - 11:15 AM
Hilton Denver City Center Lower Level 2: Denver 1-2

The use of pre-trained generative large-language models for operational assessments has increased drastically due to their capacity to generate human-like outputs. This included providing personalized feedback (e.g., Dai et al., 2023), generating assessment items (e.g., Bezirhan & von Davier, 2023), scoring open-ended responses (e.g., Lee et al., 2024), evaluating item quality, and predicting item characteristics (e.g., Liu et al., 2024). In this coordinated session, we discuss four applications of generative large-language models in operational test settings. These applications included annotating argumentative essays, providing feedback to examinees, and predicting item characteristics. This session provides an overview of different applications of generative language models as well as their limitations and challenges for operational assessment environments.

Session Organizer:
Guher Gorgun, University of Alberta

Chair:
Okan Bulut, University of Alberta

Presentations:
Leveraging Generative Language Models to Annotate Argumentative Essays
Christopher Ormerod; Kai North, George Mason University

Fine-Tuning GPT and Llama for Automatic Feedback Generation: A Comparative Study
Elisabetta Mazzullo, University of Alberta; Okan Bulut, University of Alberta

Predicting the Item Characteristics of Scenario-Based Assessment Items
Guher Gorgun, University of Alberta; Okan Bulut, University of Alberta

Predicting Mathematics Item Difficulty for Automated Item Development Workflows
Ikkyu Choi, ETS; Burcu Arslan, ETS Research Institute; Jiyun Zu, ETS

Discussant:
Ruhan Circi, American Institutes for Research

153. Can Engagement and Measurement Precision be Enhanced in NAEP Mathematics through MST?

Coordinated Paper Session
9:45 AM - 11:15 AM
Hilton Denver City Center Lower Level 2: Denver 5-6

Ensuring validity in the National Assessment of Educational Progress (NAEP), as the largest nationally representative, continuing evaluation of the condition of education in the United States, is imperative, and student engagement and measurement precision are pivotal to ensuring validity. In NAEP, multistage testing (MST) has been studied to enhance measurement precision and engagement, especially for lower-performing students. Previous research on the performance of MST in NAEP has shown that it could help improve engagement by administering items more closely aligned with students' ability level. This session presents 5 studies on student engagement and measurement precision in NAEP to inform MST design in NAEP. Study one explores if engagement differs by ability level and alignment of ability with booklet difficulty using survey responses. Next three studies examine the relationship between engagement and the alignment of ability to booklet difficulty using the percentage of students/examinees identified as disengaged using latent class analysis as a measure of engagement. Study 2 focuses on the overall relationship, excluding extended-time accommodation (ETA) sample, Study 3 examines the relationship in relation to demographic characteristics, and study 4 examines the relationship for ETA sample. Study 5 investigates the relationship between booklet difficulty and measurement precision.

Session Organizer:
Young Yee Kim, American Institutes for Research

Chair:
Markus Broer, American Institutes for Research

Presentations:
Exploring Engagement and Alignment of Ability with Block Difficulty Combination Using LCA
Young Yee Kim, American Institutes for Research

Exploring Engagement by Demographic Group, Block Difficulty, and Student Performance
Eric Muller, American Institutes for Research

Exploring the Relationships between Engagement and Booklet Difficulty for Extended-Time Accommodated Students
Xiaoying Feng, American Institutes for Research

Investigating the Relationship between Booklet Difficulty and Measurement Precision
Xiaying Zheng, American Institutes for Research

Exploring Engagement using Survey Questions by Block Difficulty Combination and Ability Group
Emma King, American Institutes for Research

Discussant:
Mark Reckase, Psychometric Solutions

154. Beyond Retrofitting: Flexibility and Comparability by Design for Equitable Assessment

Coordinated Paper Session
9:45 AM - 11:15 AM
Hilton Denver City Center Lower Level 2: Colorado CD

This session's presenters address pressing challenges at the intersection of educational measurement, assessment design, and policy for multilingual learners and students with disabilities in K-12 education. Despite advances in theory and practice, persistent performance disparities for these

populations underscore the limitations of traditional assessment approaches. This session generally will focus on thoughtful and deliberate assessment design that begins with a deep understanding of what students know and can do, while advocating for policy updates that support greater flexibility without sacrificing essential comparability. The field must shift toward more student-centered and equitable approaches to measurement, ensuring that both policies and assessments evolve to address the diverse needs of all learners, ultimately advancing inclusivity and fairness in educational outcomes. The session aims to engage measurement experts, education researchers, and policy makers in rethinking assessment frameworks and considering related policy to ensure they are flexible, equitable, and robust, ultimately promoting better educational outcomes for all students.

Session Organizer:
Edynn Sato, Sato Education Consulting LLC

Chair:
Edynn Sato, Sato Education Consulting LLC

Presentations:
Accessibility by Design: An Assessment System Case Example
Russell Swinburne Romine, University of Kansas; Meagan Karvonen; Vanessa Simmering, ATLAS, University of Kansas; Sue Bechard, ATLAS, University of Kansas

How do you know you're getting the job done?
Meagan Karvonen; Russell Swinburne Romine, University of Kansas; Amy Clark, ATLAS: University of Kansas

Advancing Equity for Multilingual Learners Who Are Blind or Have Low Vision
Laurene Christensen, WIDA at the University of Wisconsin-Madison; Trish Klein, ENTK LLC

Equitable Assessment by Design: Shifting from Standardized Tests to Mastery Metrics
Edynn Sato, Sato Education Consulting LLC; Grace Li, WIDA at the University of Wisconsin – Madison

Discussant:
Michael Russell, Boston College

155. Pushing the Utility of Assessments in Competency-Based Education

Coordinated Paper Session
9:45 AM - 11:15 AM
Hilton Denver City Center Lower Level 2: Colorado AB

This session will provide (a) an overview of competency-based education (CBE) overall and by K-12 and Higher Education, (b) indication of the role of assessments in CBE, and (c) suggestions what we as a community of researchers and measurement professionals need to do to advance our knowledge and improve our practice. Assessments in competency-education programs serve formative function providing an assessment of mastery and providing actionable feedback during the learning program and then a summative function certifying mastery of competencies which ultimately define the outcomes of the experience. Three presentations will provide an overview of CBE in general, CBE in higher education, and then CBE in K-12. The fourth presentation will discuss CBE's needs and the various research topics that researchers and practitioners have indicated are needed. Finally, a discussant with much experience in CBE as an applied researcher and practitioner will offer comments and suggestions for ensure quality.

Session Organizer:
Thanos Patelis, University of Kansas

Chair:
Thanos Patelis, University of Kansas

Presentations:
An Overview of Competency-Based Education
Ryan Specht-Boardman, Comptency-Based Education Network

State of Competency-Based Education in Higher Education and Challenges
Jessica Mason, American Institutes for Research

Competency-Based Education in K-12
Guangming Ling, ETS

Challenges, Needs, and Call for Research
Thanos Patelis, University of Kansas

Discussant:
Carla Evans, Center for Assessment

156. Cutscore Estimation Via Two-Stage Analysis: Implications for Psychometric Research and Practice

Coordinated Paper Session
9:45 AM - 11:15 AM
Hilton Denver City Center Lower Level 2: Colorado IJ

The papers of this session explore a novel approach to estimating cutscores in psychometric models, focusing on the integration of Item Response Theory (IRT) and Diagnostic Classification Models (DCMs). The session will feature five papers presenting different methodological innovations and practical applications of the two-stage estimation procedure.

The first paper introduces a maximum likelihood-based algorithm to estimate cutscores, linking IRT parameters with DCM classifications. The second paper extends this approach by incorporating Bayesian estimation, allowing for the incorporation of prior information into cutscore determination. The third paper discusses how to create substantively meaningful cutscores, essential for communicating psychometric results to non-experts. The fourth paper compares item selection strategies in computerized adaptive testing (CAT) using IRT and DCM frameworks (the latter being derived from the two-stage procedure), evaluating efficiency and accuracy. The final paper explores the application of the two-stage estimation approach in standard setting procedures, offering a novel alternative to traditional methods. This symposium will provide attendees with a comprehensive understanding of how advanced cutscore estimation techniques can enhance the accuracy and interpretability of psychometric assessments.

Session Organizer:
Jonathan Templin, University of Iowa

Presentations:
Maximum Likelihood Estimation of Scale Score Cutpoints via a Two Stage Mixture Model Likelihood
Alfonso Martinez, Fordham University

Bayesian Estimation of Scale Score Cutpoints Using a Two-Stage Mixture Model Likelihood
Jonathan Templin, University of Iowa

Creating Scale Score Classifications with Substantively Meaningful Cutpoints
Sergio Haab, The University of Iowa

Comparing Item Selection in CAT: DCM Shannon Entropy and IRT Maximum Information
Ae Kyong Jung, University of Iowa

Using the Two-Stage Estimation Approach for Cutscore Estimation in Standard Setting Procedures
Ahmed Bediwy, The University of Iowa

Discussant:
Jihong Zhang, University of Arkansas at Fayetteville

157. Electronic Board Session 6

Electronic Board Session
9:45 AM - 11:15 AM
Hilton Denver City Center Lower Level 2: Colorado GH

Presentations:

1. **Robustness of Longitudinal Diagnostic Classification Models to Model Misfit and Q-Matrix Misspecifications**
   Hyunjee Oh, Teachers College, Columbia University; Chia-Yi Chiu, Teachers College, Columbia University
   The robustness of the longitudinal DCMs is examined in the study by evaluating the impacts of model misfit and Q-matrix misspecifications on the models. The findings of this research will help researchers and practitioners in analyzing longitudinal CD data with more reliable and interpretable outcomes.

2. **DINA for Big Data Applications: Efficient Attribute Hierarchy Learning and Parameter Prediction**
   Derek Fay, Pearson; Maria Lentini, Pearson; George Wong, Pearson; Haiyan Wang, Pearson; Peter Sabatini, Pearson; Keith Augustsson, Pearson; Zahra Sajidin, Pearson
   This research enhances the scalability and utility of the DINA model in large-scale educational settings by developing efficient methods for learning attribute hierarchies and item parameters. Early results show promise in improving personalized learning experiences and supporting data-driven instructional decisions.

3. **Evaluating Attribute Importance of a Sequential CDM Using Machine Learning Methods**
   Qingzhou Shi; Qiwei He, Georgetown University; Wenchao Ma, University of Minnesota; Mehdi Rajeb
   This study integrates machine learning techniques with Cognitive Diagnosis Models (CDMs) to enhance Q-matrix validation. Using PIAAC's Problem Solving in Technology-Rich Environments (PSTRE) data, Random Forest and Logistic Regression algorithms identified key attributes contributing to proficiency classification. Findings demonstrate the effectiveness of machine learning in optimizing cognitive diagnostic assessments.

4. **Assessing the Performance of Wald Test in Detecting Item Parameter Drift in Longitudinal Cognitive Diagnosis Model**
   Benjamin Kweku Lugu, University of Alabama; Yue Yin, University of Illinois at Chicago; Ying Chen, University of Illinois System – IWERC; Wenchao Ma, University of Minnesota
   This study examines the performance of the Wald test for detecting Item Parameter Drift (IPD) based on the longitudinal LDMs. The findings show that the Wald test provides a rich source of information for detecting differences in item parameters, ensuring that items measuring the same constructs are fair over time.

5. **Examining Methods for Handling Missing Data via Exploratory Graph Analysis**
   Tong Wu; Hyeonjoo Oh, Riverside Insights ; David Dailey; Youngmi Cho, Riverside Insights
   This study evaluates the impact of five methods for handling missing data using Exploratory Graph Analysis, an innovative method for estimating factor structures. Through simulations, it assesses the relative performance of the five methods to guide practitioners in selecting the most appropriate approach.

6. **Misspecification Sensitivity of Fit Indices with Categorical Data comparing ESEM and CFA**
   Hahyeong Kim, University of Illinois at Urbana-Champaign; Jiyoon Lee, Ewha Womans University
   This study aims to the sensitivity of goodness-of-fit indices to model misspecification in CFA versus ESEM using ordered categorical data. Through Monte Carlo simulations, the impact of cross-loadings, sample size, and threshold symmetry on model fit indices is examined, offering insights into the detection of misspecification in educational measurement contexts.

7. **Does Handwriting Outshine Typing in Second Language Listening?: Insights from DIF and Beyond**
   Jieun Kim, University of Hawaii at Manoa
   This study investigates the effects of note-taking modes—handwriting and typing—on second language learners' listening performance and note quality. While overall test scores showed no significant difference, variations were observed at the item level and in note content. These insights are crucial for refining note-taking policies in language assessments.

8. **The Effect of the Next-Generation Transition on Test Validity**
   Lucia Liu, Ascend Learning; Xuechun Zhou
   The research supports that the next-generation assessments offer greater predictive validity, with this effect more pronounced for low-performing students. One reason for this could be that low-performing student benefit more from partial credit scoring, which allows the NGN assessments to more accurately reflect their actual abilities.

9. **Performance of Six BERT Models in Detecting Essays Produced by ChatGPT**
   Selene Lee, ETS
   This research examined the performance of six BERT models in classifying whether an essay was written by a student or written/paraphrased by ChatGPT. All models had a classification accuracy of above 78% for all tasks. When all layers of the model were trained, classification accuracy reached 100% for certain tasks.

10. **Comparing Paper and Online PreACT: An Alternative Approach to Mode Comparability Studies**
    MinJeong Shin, ACT, Inc.; Jing-Ru Xu
    This study provides guidelines to support evidence the mode effect is negligible when a traditional mode comparability study is unavailable. The results indicate the mode effect in the online pilot was insignificant, allowing the use of the same raw-to-scale score conversion tables for both online and paper administrations.

11. **Examining the Consistency of SAT Score Relationships with College GPA Over Time**
    Jessica Marini, College Board; Paul Westrick, College Board; Emily Shaw, College Board
    While there is plentiful evidence demonstrating the value of SAT scores to predict first-year college GPA, there is less evidence examining how well these relationships hold through each year of college. This study examines SAT score relationships with GPA through college completion, for a national sample and key subgroups.

12. **Comparison of Rasch Pre-equating and Post-equating when Item Parameters Drift**
    Chunyan Liu, National Board of Medical Examiners; Zhongmin Cui, CFA Institute; Richard Feinberg, NBME; Raja Subhiyah
    Item parameter drift (IPD) is a concern in large-scale assessments, potentially impacting validity, fairness, and reliability of test scores. In this simulation study, we investigated several IPD conditions and corresponding effect on Rasch true-score pre-equating and post-equating. Preliminary results suggest that, when IPD exists, post-equating consistently outperformed pre-equating.

13. **Censored Confirmatory Factor Analysis for Non-Normal Likert-Type Data in Language**
    Sooyong Lee, WIDA at the University of Wisconsin-Madison; Kyoungwon Bishop; Hacer Karamese, WIDA at the University of Wisconsin-Madison; Syed Abdul Hadi
    This study examines the Generalized Tobit Estimator (GBIT) as a computationally efficient alternative to Continuous CFA and GPCM for analyzing doubly-censored data from the WIDA Alternate ACCESS test, involving 3,000 students with disabilities across four language domains. A small simulation validates GBIT's utility for large-scale educational assessments.

14. Examining Response Time Effort in the Next Generation Science Standards Computerized-Adaptive Tests
Yi-Fang Wu, Cambium Assessment, Inc.; Jiajun Xu, Cambium Assessment; Mengyao Cui, Cambium Assessment, Inc.
We investigate students' response time effort in taking NGSS CAT by identifying rapid-guessing behavior. We focus on key item attributes (science disciplines, interaction types, difficulty, discrimination, and position) and student characteristics (reporting group membership, test performance, testing time, and person fit). Methods to identify RBG and practical implications are discussed.

158. Causal Models in Educational Measurement

Individual Paper Session
9:45 AM - 11:15 AM
Hilton Denver City Center Lower Level 2: Denver 4

Chair:
Anh Thu Le, University of Notre Dame

Presentations:

Multivariate Regression Discontinuity Analysis with Multidimensional Measurement Models
Youjin Sung, University of Maryland; Yang Liu, University of Maryland; Youngjin Han, University of Maryland; Muwon Kwon, University of Maryland; Ji Seung Yang, University of Maryland
Recently, methods integrating latent variables into traditional regression discontinuity (RD) analysis have been proposed, enabling the examination of heterogeneity and generalizability of the local average treatment effect. Our study extends this latent variable approach to the multivariate RD analysis where multiple assignment variables and cutoffs are used for treatment assignment.

Pseudo Maximum Likelihood Estimation of Regression Discontinuity Models with Latent Variables
Youngjin Han, University of Maryland; Youjin Sung, University of Maryland; Muwon Kwon, University of Maryland; Ji Seung Yang, University of Maryland; Yang Liu, University of Maryland
In estimation of regression discontinuity (RD) models with latent variables, re-using pre-calibrated measurement parameters can be beneficial. This study adopts pseudo maximum likelihood (PML) estimation for such model fitting scheme. PML estimation provides effective solutions for practical scenarios and ensures efficient estimation of parameters that constitute the average treatment effect.

Navigating Non-Orthogonality: Exploring Latent Mean Differences in Unbalanced Conditions
Jay Jeffries, University of Nebraska-Lincoln; James Bovaird, University of Nebraska-Lincoln
This study explores the effect of unbalanced group sizes on the accuracy of latent mean difference estimates when using orthogonal contrasts in measurement models. Monte Carlo simulations reveal that large group disproportionality seriously distorts latent mean differences by hindering statistical power and parameter efficiency while increasing error rates and bias.

Causal Mediation and Functional Outcome Analysis with Process Data for Program Evaluation
Youmi Suk, Teachers College, Columbia University; Chan Park, University of Illinois Urbana-Champaign
This study introduces causal mediation and functional outcome analysis to evaluate educational programs from response process data. We define various causal effects involving a functional outcome, and provide identification and estimation methods by adapting the mediation framework. We demonstrate our framework in the National Assessment of Educational Progress study. This study.

Discussant:
Joseph Kush, James Madison University

159. Innovations in Test Assembly

Individual Paper Session
9:45 AM - 11:15 AM
Hilton Denver City Center Lower Level 1: Penrose 2

Chair:
Chansoon (Danielle) Lee

Presentations:

Improving Ability Estimation Accuracy for Automated Item Generated Forms under Multistage Testing
Stella Kim, University of North Carolina Charlotte; Won-Chan Lee, University of Iowa
This study examines how to handle item clones generated under an automated item generation (AIG) system and evaluates the impact on ability estimation accuracy in the context of multistage testing. Additionally, the study proposes a novel method that adjusts examinees' theta estimates by accounting for random fluctuations in item difficulty.

Comparison of module assembly in multi-stage adaptive testing of a large-scale assessment
Sukkeun Im, Pearson
The simulation study examines the module assembly in large-scale assessment multistage test (MST) by comparing a traditional preassembly approach with different conditions of an adaptive module approach. Initial results show evidence of comparable measurement precision between different approaches and conditions. Further research is needed before the implementation in large-scale assessment.

Use of text embeddings in test assembly
Anna Topczewski, Law School Admissions Council; Jeremy Sakovich, Law School Admission Council; David Kary, Law School Admission Council; Dmitry Belov, Law School Admission Council
We have used text embeddings to identify pairs of semantically similar items that should not be together on the same test (enemy items) and to classify items by a topic. This allows for the optimization of test assembly by constructing tests with balanced topics and without enemies.

Optimizing Enemy Item Removal with Network Analysis in Automated Test Assembly
Charles Fisk, NBOME
This study introduces a new method utilizing network analysis to optimize enemy item removal in automated test assembly (ATA). This new method minimizes item loss across item pools, maintains balanced item retention, and improves ATA efficiency, particularly under high enemy severity conditions.

Discussant:
Paulius Satkus

160. Generalizability Theory and Structural Equation Modeling Individual Paper Session

9:45 AM - 11:15 AM
Hilton Denver City Center Lower Level 2: Denver 3

Chair:
Ernest Amoateng, Western Michigan University

Presentations:

Validating the Parenting Stress Index for Latinx Parents of Children with Disabilities
Hyeri Hong, California State University, Fresno; Kristina Rios, California State University, Fresno
We investigated item wording effects of the Parenting Stress Index-Short Form for Latinx parents of children with intellectual and developmental disabilities (IDD) within structural equation modeling-based generalizability theory frameworks. This study will suggest that culturally insensitive items need to be refined to appropriately reflect the stress faced by Latinx parents.

Using G-theory-based Structural Equation Models to Account for Item Wording Effects
Walter Vispoel, The University of Iowa; Hyeri Hong, California State University, Fresno
We used SEM-based G-theory analyses to partition total-score variance into trait, state, method, item-wording, and random-error effects using data from a popular multidimensional self-concept inventory. Results underscored the importance of separating item wording effects from other sources of variance when evaluating the reliability and validity of scores.

Estimating Score Dependability in Congeneric and Essential Tau-Equivalent Generalizability Theory-based SEM Designs
Hyeryung Lee; Walter Vispoel, The University of Iowa
We use structural equation models to estimate G-theory-based absolute-error, relative-error, dependability, and generalizability coefficients for essential tau-equivalent (ETE) and congeneric (CON) models. Findings highlighted the efficacy of our approach across univariate, multivariate, and bifactor designs with CON models producing better model fits, higher score consistency coefficients, and lower measurement error.

Applying Multivariate Generalizability Theory to Personality Assessments within Structural Equation Modeling Frameworks
Hyeri Hong, California State University, Fresno; Walter Vispoel, The University of Iowa
Despite having such advantages as measuring score consistency and identifying various sources of measurement error, multivariate generalizability theory is rarely applied to personality assessments. In this study, we demonstrate how to use multivariate generalizability theory with such assessments to expand and complement univariate procedures within structural equation modeling frameworks.

Discussant:
Kimberly Colvin, University at Albany, SUNY

161. Bridging Theory and Practice: Fostering Innovation in Diagnostic Measurement to Enhance Education (Organized Session)

Organized Discussion
9:45 AM - 11:15 AM
Hilton Denver City Center Lower Level 1: Penrose 1

This organized discussion brings together leading voices from academia and industry to explore the latest research and trends in diagnostic measurement (DM) that are both theoretically robust and practically relevant in today's changing educational settings. The session will showcase cutting-edge models and techniques, such as learning progressions, Diagnostic Classification Models (DCMs), artificial intelligence (AI), and Structured Mixture Item Response Theory (SMIRT) models, all aligned with emerging educational priorities. These include supporting empowered learning in the formative

assessment process, improving the precision and insights of diagnostic assessments, and enhancing instructional decision-making through validated learning progressions and diagnostic feedback at multiple levels. The discussion will emphasize actionable solutions, highlight practical challenges and successes, and encourage audience engagement in exploring how DM can be most effectively utilized. By integrating diverse perspectives from both researchers and practitioners, this session aims to advance the understanding and application of DM to better serve society's educational needs, with a focus on practical applications and societal implications. The interactive format is designed to ensure that insights shared are directly applicable, fostering a collaborative vision for the future of DM.

Chair:
Shiyu Wang, University of Georgia

Moderator:
Yon Soo Suh, NWEA within HMH

Presenters:
Jake Thompson
Laine Bradshaw, Pearson
Richard Golden
Minjeong Jeon, UCLA
Minho Lee, University of California - Los Angeles

162. Psychometrically-driven Product Development in the Age of AI
Organized Discussion
9:45 AM-11:15 AM
Hilton Denver City Center Lower Level 2: Colorado EF

Building products and capabilities in educational measurement that are AI-driven or AI-enhanced requires multi-disciplinary teams and perspectives that can include product managers, content developers, software engineers, data scientists, UI/UX engineers, hand-scoring experts, learning scientists, and instructional designers. Psychometricians are not always trained to work in these multi-disciplinary, product development contexts, and may even shy away from them as too complex, too messy, and not "measurement." Worse, psychometricians can be seen as irrelevant or roadblocks to the product development process. It is, however, critical that they play a substantial role in the development and validation; otherwise, the product design may lack a coherent theory of action, design that supports the theory, and research that seeks to provide evidence that supports appropriate use. In this session, we discuss how to participate, and lead, this process from four very different perspectives and use cases. Each speaker in this session plays a different role in defining what a product is, who it is for, and how it should be used. We will focus on actionable guidance on how to influence the product development process, either as a leader, subject matter expert, proposer of new products, or reviewer.

Chair:
Harold Doran, HumRRO

Moderator:
Harold Doran, HumRRO

Presenters:
Susan Lottridge, Cambium Assessment
Jennifer Dunn, Pearson
Alina von Davier, Duolingo
Kara McWilliams, National Council of Bar Examiners

163. Socioculturally Responsive Assessment: Implications for Theory, Measurement, and Systems-Level Policy

Coordinated Paper Session
11:30 AM - 1:00 PM
Hilton Denver City Center Lower Level 2: Colorado CD

Socioculturally responsive assessment (SCRA) is an emerging concept within educational measurement that was synthesized from multiple literatures, including those on the assessment and instruction of minoritized students and the learning sciences. SCRA's underlying premise is that designing assessment in ways that account for examinees' social, cultural, and other relevant characteristics should foster more accurate representations of what students know and can do. Such design, however, poses significant challenges. The purpose of this session is to preview the latest thinking in this emerging area, to appear in an upcoming open-access book, Socioculturally Responsive Assessment: Implications for Theory, Measurement, and Systems-Level Policy (Bennett, Darling-Hammond, & Badrinarayan, in press). One of the editors (Badrinarayan) will give a brief overview, followed by presentations from four of the chapter authors. This sampling will touch on several of the book's key ideas, including framing of constructs within SCRA, measurement challenges and potential solutions, and empirical research results.

Session Organizer:
Randy Bennett, ETS

Chair:
Randy Bennett, ETS

Presentations:
Socioculturally Responsive Assessment: Volume Overview
Aneesha Badrinarayan, Education First

Socioculturally Responsive Large-Scale Assessment: The Case of the 2026 NAEP Reading Framework
Allison Skerrett, University of Texas at Austin

Linking Responsive Assessments: Challenges and Possibilities
Tim Moses, Buros Center for Testing

An Evidentiary-Reasoning Lens for Socioculturally Responsive Assessment
Robert Mislevy, University of Maryland

Group Differences Across Scenario-Based Reading Assessments: Examining the Effects of Culturally Relevant Test Content
Zuowei Wang, ETS

164. Understanding AI and Human Ratings: Complementary Roles of AI and Human Raters

Coordinated Paper Session
11:30 AM - 1:00 PM
Hilton Denver City Center Lower Level 2: Denver 3

With the advances in Large Language Models (LLMs) and Generative AI or Chatbots, both have been explored for automated scoring in education assessments. While LLMs such as BERT, DeBERTa, RoBERTa, and ELECTRA are often trained for automated scoring in high-stakes testing programs, Chatbots such as ChatGPT, Claude, and Gemini can be prompted for automated scoring for low-stakes assessments. The session presents studies to help understand how AI ratings are compared with human ratings. More specifically, studies in this session critically assess the reliability of AI and human ratings, estimate the AI rater effects compared with human rater effects, enhance scoring accuracy through hybrid human-AI systems, and compare scoring rationales between AI and human raters for non-English essays. These studies employ diverse methodologies, including multivariate

generalizability theory analyses, Many-Facet Rasch modeling, and machine learning techniques. They explore applications ranging from AP Chinese writing tasks to science teachers' pedagogical content knowledge assessments. Studies in this session critically examine the strengths and limitations of AI-based scoring demonstrating where AI excels, where human expertise remains crucial, and propose strategies for leveraging the complementary strengths of human and AI raters to enhance the validity, reliability, and efficiency of automated scoring in educational assessment programs.

Session Organizer:
Hong Jiao, University of Maryland

Chairs:
Hong Jiao, University of Maryland
Won-Chan Lee, University of Iowa

Presentations:
Exploring LLM Autoscoring Reliability in Large-Scale Writing Assessments Using Generalizability Theory
Dan Song, The University of Iowa; Won-Chan Lee, University of Iowa; Hong Jiao, University of Maryland

Estimating Human and AI Rater Effects Using the Many-Facet Rasch Model
Hong Jiao, University of Maryland; Dan Song, The University of Iowa; Won-Chan Lee, University of Iowa

Human Raters and Automated Scores: Hybrid Scoring for the Best of Both Worlds
Alexander Kwako, Cambium Assessment; Christopher Ormerod; Susan Lottridge, Cambium Assessment

Evaluating Rationales: A Comparative Study of LLMs and Human Raters in Scoring Language Learners' Writing
Jodi Casabianca, ETS; Dan Song, The University of Iowa; Magdalen Beiting-Parrish, Federation of American Scientists; Sharon Slater, ETS; Hong Jiao, University of Maryland

Examining the LLMs' Rater Reliability for Contextual Constructed Response Assessments
Shuchen Guo, University of Georgia; Xiaoming Zhai, University of Georgia

Discussant:
Mark Shermis, Performance Assessment Analytics, LLC

165. Using NAEP Data to Quantify and Address Educational Inequities

Coordinated Paper Session
11:30 AM - 1:00 PM
Hilton Denver City Center Lower Level 2: Denver 4

Educational disparities based on demographic factors have been a persistent trend in the United States, revealing the ongoing challenges in ensuring equitable educational opportunities, particularly for those students from marginalized groups. This coordinated session examines these inequities using data from the NAEP Mathematics Assessment and Technology Landscape Survey through five studies, using data about students and their schools. As the largest nationally representative assessment, NAEP offers critical insights for shaping public policy and identifying inequality. Four studies explore non-cognitive, cognitive, and behavioral factors in mathematics performance: one reveals racial and gender gaps but improved outcomes for English learners and students with disabilities, while another shows how effective digital tool use benefits minoritized and low-income students. Two studies on Puerto Rican students highlight a sharp decline in foundational math skills and emphasize the role of teacher support and resources in improving confidence and achievement during COVID-19. A final study examines the transition to school-based devices for NAEP assessments, finding that while most schools meet network requirements, rural and marginalized

communities face ongoing technology access challenges. In the service of society, this session underscores how NAEP data can quantify educational inequities and inform interventions to promote equitable access to technology and instruction.

Session Organizer:
Fernando Mena, University of Massachusetts Amherst

Chair:
Fernando Mena, University of Massachusetts Amherst

Presentations:
Identifying Relevant Mathematics Engagement Indicators Using NAEP Process Data
Antranik Kirakosian, Washington State University

The Transition to School Devices for NAEP
Maya Robertson, Grambling State University

Comparing NAEP Mathematics Items Performance Between 2019 and 2022 in Puerto Rican Students
Fernando Mena, University of Massachusetts Amherst

Exploring the Association of Non-Cognitive Factors with NAEP Mathematics Scores for Puerto Rican Students
Sandra Cruz, DEMRE - Universidad de Chile

Discussant:
Luis Saldivia, ETS

166. Research for Practical Issues and Solutions in Computerized Multistage Testing (Routledge Book)

Coordinated Paper Session
11:30 AM - 1:00 PM
Hilton Denver City Center Lower Level 1: Penrose 2

Artificial Intelligence(AI) has received a lot of attention, but much of the focus is on large language models (LLM). Often overlooked is that adaptive testing is a form of AI, leading AI research since the 1960s, such as the algorithms for merging adaptive learning and adaptive testing. Multistage testing(MST) provides valuable information about individuals, enabling decisions such as hiring, professional skills certification, university admissions, medical diagnosis, or determining the next teaching step for a 4th-grade math student. We aim for these tests to provide accurate information efficiently. MST maximizes the information gained per minute of testing. By administering items that match the K-12 student's level, we can effectively determine their mastery level, reducing time away from actual learning, increasing test security, and enhancing engagement by tailoring difficulty to the examinee's ability.

This symposium showcases recent research and applications in MST, marking the debut of the new book Research on Practical Issues and Solutions in Computerized Multistage Testing (2024). It provides the latest research on various practical considerations, methodological approaches, and solutions for implementing MST in operational applications, such as the 2024 Scholastic Aptitude Test(SAT). The presentations are authored by renowned psychometricians recognized globally for their expertise in adaptive and multistage testing.

Session Organizer:
Duanli Yan, ETS

Chair:
Duanli Yan, ETS

Presentations:
AI's Impact on Adaptive Testing
Alina von Davier, Duolingo

History and Introduction of Adaptive and Multistage Testing
Nathan Thompson, Assess Systems Incor; Duanli Yan, ETS; David Weiss, University of Minnesota

Designing MST to Meet Accuracy and Efficiency Goals
Mark Reckase, Psychometric Solutions; Unhee Ju, Cognia

Innovative Approaches for MST Routing
Kyung (Chris) Han, GMAC

Development and Applications of Probability-Based Classification for MST
Victoria Song, Fordham University; Duanli Yan, ETS

Discussant:
Oliver Zhang, College Board

167. Dynamic Measurement of Student Proficiency for Continuous Assessment

Coordinated Paper Session
11:30 AM - 1:00 PM
Hilton Denver City Center Lower Level 2: Colorado AB

In today's classrooms, students interact with technology – digital assessments, digital instruction, digital learning games – on a near daily basis. Many of these activities present opportunities for measurement but do not, by themselves, contain enough measurement information to produce reliable and valid scores. To provide continuous insights into student proficiency throughout the school year requires a dynamic measurement model capable of using information across many activities where learning happens between each activity.

The papers in this session explore three different approaches for using weekly measurement data from i-Ready, a digital assessment and instruction system, to dynamically measure student math proficiency throughout the school year. The models discussed range from Bayesian extensions of Item Response Theory (Glicko-IRT, Empirical Bayes) to Machine Learning approaches (XGBoost, Deep Learning). Each paper leverages the same i-Ready dataset and attempts to achieve the same goal. The models are compared and evaluated using a common framework, described in the last paper. Validity evidence is evaluated by comparing dynamic measurement model scores and placement levels to those achieved on the i-Ready Diagnostic interim assessment. Finally, the models are also evaluated on their ease of implementation and explainability to educators.

Session Organizer:
Logan Rome, Curriculum Associates

Chair:
Logan Rome, Curriculum Associates

Presentations:
Combining IRT with Glicko for the Purpose of Dynamic Measurement
Ted Daisher; Erin Banjanovic ; Logan Rome, Curriculum Associates

An Application of XGBoost and Deep Learning Models to Predict Student Performance
Dean Serrentino, Curriculum Associates; Keith Holliday, Curriculum Associates; Mihye Choi, Curriculum Associates

A Self-Contained Empirical Bayes Approach to Weekly Scoring
Sanford Student, University of Delaware; Derek Briggs, University of Colorado – Boulder

A Review and Evaluation of Three Approaches to Dynamic Measurement
Erin Banjanovic; Logan Rome, Curriculum Associates; Ted Daisher

Discussant:
Ye Tong, National Board of Medical Examiners

168. Graduate Student eBoard Session 5

Graduate Student Electronic Board Session
11:30 AM - 1:00 PM
Hilton Denver City Center Lower Level 2: Colorado GH

Presentations:

1. **Bridging AI and Educational Measurement: A Systematic Review and Curriculum Reform**
   Henry Makinde; Hope Adegoke, University of North Carolina
   This systematic review explores the integration of AI in educational measurement in light of recent advances with LLMs and emerging psychometrics job market needs for AI-based skills. We call for curriculum reform in graduate-level studies to meet these emerging needs and to prepare future professionals in AI-driven educational assessment professions.

2. **Incorporating Sparsity into Bayesian Stacking Procedures**
   Kjorte Harra, University of Wisconsin-Madison; David Kaplan, University of Wisconsin - Madison
   Bayesian stacking procedures combine multiple models to optimize prediction. We examine whether adding regularization through sparsity-inducing priors in stacking improves out-of-sample predictions through decreasing variance, as previous work suggests. We find that inducing sparsity in member models provides no benefit to the stack despite the stronger performance of member models.

3. **Supporting Educational Outcomes through Gamified Computational Thinking Assessments**
   Tiantian Hou, Beijing Normal University; Jian Li, Beijing Normal University
   This study investigates the impact of gamified computational thinking (CT) assessments in middle schools. The finding demonstrates that gamified elements increase student motivation and participation, while maintaining assessment integrity. These results provide empirical support for the effectiveness of gamified CT assessments as a means to inform and enhance educational practices.

4. **Using a Mixture Speed and Accuracy Model to Diagnose Oral Reading Fluency**
   Paul Foster, Southern Methodist University; Yusuf Kara, University of Miami; Akihito Kamata
   Identifying latent classes of students in oral reading data may afford diagnostic analysis to guide instruction. We demonstrate a mixture model fitted to passage-level time and accuracy data. Implications and future extensions of the model are discussed with preliminary data analysis.

5. **Student Perception Towards Different Performance-based Assessment Approaches**
   Juan Arboleda; R.R. Meijer, University of Groningen; A.S.M Niessen, University of Groningen; M. Neumann, Free University of Amsterdam
   This study investigated student perceptions of different degrees and components of standardization in performance-based assessment in higher education. Findings suggest that high degrees of standardization and absolute judgment lead to more positive perceptions than low standardization and comparative judgment. These insights have implications for assessment design in higher education.

6. **An initial investigation of the internal consistency reliability of language learning progressions.**
   Preston Botter; Ève Ryan, University of Alaska Fairbanks; Alison Bailey, UCLA; Dubravka Svetina Valdivia, Indiana University
   This study estimates internal consistency reliability for tasks rated on domains expressed as language learning progressions (LLPs) using Cronbach's alpha variants. Most LLPs exceed the 0.74

reliability threshold, aligning with second language research benchmarks, indicating LLPs' strong reliability for instructional and assessment use in classrooms.

7. **Work in Progress: Optimizing Q-matrix Using Association Rule for consecutive assessments with similar attributes.**
   Amirreza Mehrabi, Purdue University; Jason Morphew, Purdue University
   We explore skill mastery in educational assessments using unsupervised learning and association rule mining for consecutive assessments with similar attributes. Analyzing data from the Force Concept Inventory and Force and Motion Conceptual Evaluation, Study identifies co-occurring mastery subsets and evolving profiles, informing revisions of attribute profiles through DINA and ML.

8. **Improving IRT Subscores Using Methods from DCM**
   Adam Coates, University of Bath
   Subscores are often highly correlated anad so provide little valuable feedback. Methods developed for Diagnostic Classificaiton Models can be transferred to the Item Response Theory framework to potentially improve the distictiveness and reliability of subscores. This study uses TIMSS data to explore how this approach can improve methods for producing subscores.

9. **How Does Rapid Guessing Distort Estimates in Longitudinal Diagnostic Models?**
   Nicolas Mireles, James Madison University; Yu Bao, James Madison University
   This study examines the impact of rapid guessing on longitudinal Diagnostic Classification Models. An empirical analysis of 1,634 college students' test data was analyzed, followed by a simulation study manipulating various factors for a longitudinal diagnostic assessment. Results indicate rapid guessing behaviors change interpretations of population and individual progress.

10. **Factor Structure and Invariance of Learning Outcomes Scale in Norwegian Student Survey**
    Lucy Gitiria, UMass Amherst
    Scores from this learning outcomes measure are key quality indicators in higher education. For validation purposes, its factor structure was investigated, and measurement invariance was tested across four diverse study programs. Multigroup confirmatory factor analysis on data from 7,488 students revealed the modified structure was not fully invariant across programs.

11. **Comparing Methods for Detecting Measurement Invariance Across Multiple Groups**
    Haoyang Yu, University of Kansas; Sean Joo, University of Kansas
    This study compares RMSD-based DIF detection methods with the conventional methods such as generalized Mantel-Haenszel, generalized logistic regression, and Lord's chi-square tests. A simulation study evaluates their performance in detecting DIF for multiple groups. Empirical results of the DIF methods will be demonstrated using PISA 2022 data.

12. **Beyond the Surface: A Comparative Study of MNLFA and Traditional DIF Analysis**
    Autumn Wild; Joseph Kush, James Madison University
    Items with evidence of differential item functioning (DIF) pose a serious threat to validity. Moderated Nonlinear Factor Analysis (MNLFA) models provide enhanced flexibility in examining DIF compared to traditional methods. We present desirable results, advocating for the use of the MNLFA model over traditional DIF analytic approaches.

13. Youth Experiences Survey (YES!): An Item and Differential Response Function Analysis
Angie Williams
The Youth Experience Survey (YES) 2.0 is a self-report scale for adolescents measuring their developmental experiences in organized youth activities. This paper examines the dimensionality of the scale and examines the item parameters to determine function and potential DIF. Results inform scale revision supporting development of a short form scale.

14. A Comprehensive Comparison of Partially Confirmatory Methods in SEM
Yifan Zhang, The University of Hong Kong; Jinsong Chen, The University of Hong Kong
The partially confirmatory issue is an important challenge to balance exploratory and confirmatory in SEM. Well-known approaches like ESEM, BSEM, Regularized SEM and Bayesian regularization have been widely used to achieve this. This study will provide a comprehensive evaluation and comparison across these methods, with different structures and estimate conditions.

15. Simulation-based Power Analysis for DIF/DTF in IRT Graded Response Models
Austin Wyman, University of Notre Dame; Zhiyong Zhang, University of Notre Dame
WebPower is a free collection of tools for conducting statistical power analysis online. A new Monte Carlo method is proposed and implemented, which allows users to conduct Monte Carlo Simulation-based power analysis for DTF and DIF for IRT Graded Response Models. Anchors, DIF effect size, and uniformity may be specified.

16. The Effects of Gender-Biased Academic Achievement and Student-Perceived Teacher Support
Christina O'Grady
This study compared the relationship between high school students' perceptions of teacher support in mathematics, gender, and academic success. The results found students' perceptions did not impact mathematics achievement scores. However, there was a trend in students' perceptions of gender and its impact on their lives, showing a generational transition.

169. Expanding Automated Scoring Validity & Generalizability using AI

Individual Paper Session
11:30 AM - 1:00 PM
Hilton Denver City Center Lower Level 2: Denver 1-2

Chair:
Kelsey Nason, James Madison University

Presentations:

Applying Cross-Prompt (Generic) Automated Scoring Models to a Large-Scale Writing Assessment
Scott Wood, ACT, Inc.
Cross-prompt automated essay scoring (AES) models are trained using data from multiple comparable prompts and applied to essays from other comparable prompts in the bank. Cross-prompt models from a large-scale writing assessment show that the models performed as well as hand scorers across a series of agreement metrics.

Validation Beyond Agreement: Investigating Human and Automated Scorer Evaluation of Textual Changes
Sarah Hughes
This paper seeks to expand the construct validity evidence for assessments that use automated scoring by investigating additional forms of validation evidence beyond human/machine agreement. This study uses expert human judgement alongside eXplainable AI (XAI) techniques to identify influential factors in essay scoring decisions and measure their impact.

Comparison of Human Rater and Automatic Scoring on Students' Ability Estimation
Zhen Wang, Hkeaa
The purpose of this research study is to examine examinees' ability estimation using IRT-based rater model. First, the examinees' ability estimates based on human raters and automatic scoring are compared; Second, the scoring results based on the different scoring designs at student level, prompt level and test level are examined.

Assessing the Representativeness of LLM-Generated Item Responses Using Latent Class Analysis
Yining Lu; Yuan Huang, New York University; Klint Kanopka, New York University
We evaluate LLM's ability to replicate human responses to items from the General Social Survey. Assessing multiple models, we find variation in their ability to predict individual responses but no ability to reproduce latent class structure or response distributions. We advise caution to researchers using LLMs to generate item responses.

Discussant:
Diego Zapata-Rivera, ETS

170. DCM for Longitudinal Data

Individual Paper Session
11:30 AM - 1:00 PM
Hilton Denver City Center Lower Level 2: Denver 5-6

Chair:
Youn Seon Lim, University of Cincinnati

Presentations:
Designing Learning Intervention Studies: Identifiability of heterogeneous hidden Markov models
Steven Culpepper, University of Illinois at Urbana-Champaign; Ying Liu, University of Illinois at Urbana-Champaign
Hidden Markov models (HMMs) are a popular framework for tracking student learning. HMMs provide a powerful framework for understanding student learning, however, the classical framework requires assumptions that are rarely met in the educational context. We discuss different types of identifiable assessment designs and the implications for practice.

Evaluating learning progression hierarchies with diagnostic models
Auburn Jimenez, University of Kansas; Jake Thompson; Brooke Nash, University of Kansas
To have confidence in inferences drawn from learning progressions, we need methods for evaluating their accuracy. In this paper, we describe how diagnostic models were used to evaluate the structure of learning progressions underlying an innovative instructionally embedded assessment system.

A longitudinal diagnostic facet status model: Tracking the learning trajectory
He Ren, University of Washington; Chun Wang, University of Washington
The diagnostic facet status model (DFSM) provides instructional-useful information on students' learning status on goal facets and misconceptions. This study extends DFSM to a longitudinal context, tracking students' facet profiles dynamically. A regularized expectation-maximization algorithm is proposed to uncover the longitudinal learning patterns. The model is evaluated by simulation studies.

Discussant:
Madeline Schellman, Pearson

171. Innovations in Technology-Based Assessment

Individual Paper Session
11:30 AM - 1:00 PM
Hilton Denver City Center Lower Level 2: Colorado IJ

Chair:
Kaiwen Man, University of Alabama

Presentations:

Routing Strategies for a Multiple Subscale Multistage Test
Ru Lu, Educational Testing Service; Paul Jewsbury, ETS; Shuhong Li, Educational Testing Service
This study compared three multistage testing (MST) designs for measuring two subscales: independent (baseline), intersectional and integrated routings (alternatives). Results showed that with highly correlated subscales, the alternative designs provided similar reliability and accuracy despite being shorter than the independent MSTs, suggesting they are effective options without compromising test quality.

Linking an English Language Test (GTELP Junior) to the CEFR: Embracing Foundational Teaching in the Age of AI
Minjung Kim
This study links the GTELP Junior English Proficiency Test to CEFR levels using the modified Angoff method. Through expert panel evaluations, statistical validation, and comparisons with the Bookmark method, the results show reliable score alignment across CEFR levels, enhancing the test's global relevance and supporting evidence-based language assessment standards.

A Critical Analysis of a Tool for Quantitative Ethnography
Trina Emler, University of Kansas, AAI, CLARA; Angela Murray, University of Kansas; Heather Gerker, University of Kansas
Innovative data processing and visualization tools can provide new pathways for measurement and analysis and often present themselves as inviting alternatives to traditional forms of measurement. Our research presents a targeted use case of one such tool, focusing on both the affordances and constraints for analysis of data sets.

Discussant:
William Skorupski, Data Recognition Corporation

172. The Future of K12 Assessment & Accountability: Can Recent Research inform policy changes?

Organized Discussion
11:30 AM - 1:00 PM
Hilton Denver City Center Lower Level 2: Colorado EF

The standards-based reform movement shaped the last 40 years of education, which resulted in teaching and learning improvements as well as unintended consequences, particularly for Black, Latinx, Native Indigenous, and low-income students.

Join us for a compelling discussion focused on how the next 40 years in K12 assessment and accountability are being reimagined. Attendees will hear from Education Analytics about a new model for accountability, the Equity-Aligned Analytics System (EAAS), that leverages predictive analytics to provide teachers, school administrators, students, and parents with real-time data to increase student readiness to graduate high school. In this session, The University of Kentucky and the Center for Innovation in Education will illuminate how innovation in K12 assessment and accountability can start from the ground up at the local and state level when driven by equity and community voice to inform changes in accountability policy at the federal level. Finally, Bellwether will share how four states piloted through-year assessments and the opportunities to incorporate parent voices in assessment design and implementation decisions.

Chair:
Sophia Velez, EduDream

Moderator:
Michelle Oliva, EduDream

Presenters:
Michelle Oliva, EduDream
Robert Meyer, Education Analytics & University of Wisconsin at Madison
Michelle Croft, Bellwether
Daniela Di Digiacomo, University of Kentucky

173. The Usefulness of Kane's "Validity Argument" in Modern Validity Theory
Organized Discussion
11:30 AM - 1:00 PM
Hilton Denver City Center Lower Level 1: Penrose 1

Kane's argument-based approach to validation (Kane, 1992; 2016) involved two steps: 1) articulate the claims made about test scores to facilitate interpretation and use as an argument (what Kane called IUA); and 2)evaluation of this argument based on evidence (i.e., the strength of the argument and the evidence supporting it is the validity information worth considering). For basic measures of literacy or numeracy, the supporting evidence might be simple, but for more complex claims much more evidence would be required. Kane also claimed and Sireci (2013) agreed that this argument-based approach was well aligned with the Standards for Educational and Psychological Testing (hereafter "Standards"; AERA, APA, & NCME, 2014) in that documented evidence and rationale are needed to support the proposed uses of a test. Despite this, Newton and Baird (2016) found that the debate about the evolution of validity continues. Given the recently completed but not yet published fifth edition of Educational Measurement and its chapter on validity, the pending development of the next edition of the Standards and NCME's task force proposing Foundational Competencies in Educational Measurement, (see Ho, et. al., 2024 & the call for commentary) another discussion about the usefulness of Kane's validity.

Chair:
Michael Walker, HumRRO

Moderator:
Michael Walker, HumRRO

Presenters:
Michael Walker, HumRRO
Gregory Cizek, University of North Carolina Chapel Hill
Stephen Sireci, University of Massachusetts Amherst
Catherine Welch
Mya Poe, Northeastern University

174. Advancements in Test Construction and Evaluation of Multistage Tests

Coordinated Paper Session
1:30 PM - 3:00 PM
Hilton Denver City Center Lower Level 2: Colorado CD

Computer-based testing is ubiquitous in K-12 assessment, often administered as a fixed form, but increasingly administered in computer-adaptive test (CAT) format. While item-level CATs are considered the psychometric ideal, multistage CATs (MSTs) are seen to provide most (if not all) the psychometric improvements of item-level CATs after factoring in cost, pool size, security, transparency of form development and evaluation, and fidelity to content constraints. These MST advantages have led to their increased use, resulting in the need for further development of this nascent methodology. Our session features five papers showcasing some of the most recent advancements. The first introduces a new psychometric evaluation tool: the Expected Test Information function (ETIF). The second paper presents a new versatile psychometric analytical R Shiny for quickly evaluating the scoring and routing accuracy of any MST from multiple TIF perspectives including ETIF. The third paper presents the latest research on how to best establish TIF targets for individual modules. The fourth and fifth papers together give a comprehensive view of MST construction and evaluation. One paper focuses on test construction led by content experts. The other focuses on test evaluation led by psychometricians. The iterative interaction of content and psychometrics is highlighted in both.

Session Organizer:
Louis Roussos, Cognia

Chair:
Louis Roussos, Cognia

Presentations:
Expected Value of Test Information Function and Applications to Evaluation of Multistage Tests.
Unhee Ju, Cognia; Liuhan Cai, Accenture; Louis Roussos, Cognia

Generalized TIF Evaluation R Shiny for Multistage Tests
Cheng Hua, University of Montevallo; Louis Roussos, Cognia

Establishing TIF Targets for Multistage Test Segments
Huijuan Wang; Keith Boughton, Data Recognition Corporation; Jessalyn Smith, DRC

Multistage Test Construction from Content Perspective
Katie Schmidt, College Board; Qi Qin, Cognia

Multistage Test Evaluation from Psychometric Perspective
Qi Qin, Cognia; Louis Roussos, Cognia; Unhee Ju, Cognia; Liuhan Cai, Accenture

Discussant:
Mark Reckase, Psychometric Solutions

175. Building Just Assessments: A Culturally and Community-Grounded Perspective

Coordinated Paper Session
1:30 PM - 3:00 PM
Hilton Denver City Center Lower Level 2: Colorado AB

Western education, specifically assessments, has historically marginalized Native students. Assessments built on dominant cultural values can be especially destructive when linked to high-stakes decisions (Trumbull & Nelson-Barber, 2019). However, the sustained will and community engagement that centers cultural wisdom and worldview can provide a path forward. This approach includes considering culture across all aspects of assessment development and implementation.

Outside researchers can support this work through deep collaboration and by centering the community's wisdom.

This session highlights a series of assessment projects where Native communities play a critical role in assessment development and implementation activities by focusing on culture. Two of these projects are based in Hawai'i, where a kindergarten readiness assessment and a state accountability assessment work to sustain and grow the Hawaiian language. These projects center on community and cultural validity to ensure the inclusion of community members' priorities. Another assessment project in Alaska focuses on centering Native Alaskan wisdom to inform assessment content and to show how outside researchers can best support this work.

The presentations in this session provide practical examples of a path forward for other assessment programs. They show how engaging communities can improve measurement practices to create more robust and useful assessments.

Session Organizer:
Kerry Englert, Seneca Consulting, LLC

Chair:
Kerry Englert, Seneca Consulting, LLC

Presentations:
Considering New Ideas for Culturally Sustaining Assessments
Kerry Englert, Seneca Consulting, LLC

Development of a Culturally Sustaining Kindergarten Readiness Assessment
Pohai Kukea Shultz, University of Hawaii – Manoa

Standing on Yup'ik Ground: A Process for Developing the Yup'ik Language and Culture Assessment
Gayle Sheppard Miller, Lower Kuskokwim School District, Bethel, Alaska; Vitaliy Shyyan; Edynn Sato, Sato Education Consulting LLC

Indigenous Research and Mixed Methods: Integration for Language and Culture Surveys
Edynn Sato, Sato Education Consulting LLC; Vitaliy Shyyan; Kerry Pusey, University of Pennsylvania

Discussant:
Karla Egan, EdMetric

176. Early Literacy Assessment: Developing, Validating, and Applying Measures that Support Student Learning

Coordinated Paper Session
1:30 PM - 3:00 PM
Hilton Denver City Center Lower Level 2: Denver 4

This session brings together five papers that tackle the complexities of developing, validating, and applying early literacy measures. The first evaluates three analytic methods for dyslexia screening, providing insights into how cut points and measurement approaches impact early identification efforts. The second introduces a culturally responsive assessment specifically developed for Hmong preschoolers, emphasizing the importance of engaging community perspectives in assessment development. The third assesses the reliability and validity of oral fluency tasks, demonstrating their practical utility in supporting diverse early reading skills. The fourth explores methodologies for synthesizing multiple literacy assessment results, addressing challenges in creating reliable screeners that accurately identify struggling readers. Finally, the session closes with an examination of how literacy screeners are used by states, underscoring the importance of valid uses of these scores. Together, these papers offer a comprehensive exploration of early literacy assessment, advocating for practices that are not only psychometrically sound but also culturally and contextually relevant to meet the needs of diverse learners.

Session Organizer:
Sarah Quesen, WestEd

Chair:
Sarah Quesen, WestEd

Presentations:
Comparing Analytic Approaches to Predicting Dyslexia Risk in Grades K-3
Patrick Kennedy, University of Oregon; Gina Biancarosa, University of Oregon; Brian Gearin, University of Oregon

Hmong preschooler's heterogeneity in item responses: facets of assessment design and calibration
Alisha Wackerle-Hollman, University of Minnesota; Kory Vue, University of Minnesota; Kelsey Will, University of Minnesota; Gao Vang; Lori Erickson

Evaluating Early Literacy Tasks: Insights from a Mixed-Methods Study
Catherine Oberle; Sonya Powers

Demystifying Literacy Screeners Through Synthesizing Literacy Assessment Results
Aimee Boyd; Mollie Headley, Curriculum Associates; Megan Perna, Curriculum Associates; Yu Su, Curriculum Associates

Are We On the Same Page? A Discussion of the Use and Misuse of Early Reading Assessments
Meredith Langi; Matthew Brunetti, WestEd

Discussant:
Tony Albano, University of California – Davis

177. The Fifth Edition of Educational Measurement: Specific Applications of Assessment

Coordinated Paper Session (Joint NCME/AERA Session)
1:30 PM - 3:00 PM
Hilton Denver City Center Lower Level 2: Colorado EF

Four chapters from the coming fifth edition of Educational Measurement (EM5) will be discussed in this coordinated paper session. Each of these chapters focuses on a specific application of educational assessment. The chapters are Assessment to Inform Instruction and Learning; Test-Based Accountability in K-12 Education; Assessment for Admission, Placement, and Outcomes in Higher Education; and Assessment for Licensing and Certification. A brief overview will be presented by the first author of each chapter, followed by a critique of the chapters provided by two EM5 reviewers. Audience participation will be invited following the discussants' comments. (Please note that this is one of two EM-5 focused sessions at the annual meeting.)

Session Organizer:
Susan Brookhart, Duquesne University

Chairs:
Mary Pitoniak, ETS
Linda Cook, ETS

Presentations:
Assessment to Inform Teaching and Learning
Susan Brookhart, Duquesne University; Charles DePascale, Psychometric Confections, LLC

Test-Based Accountability in K–12 Education
Andrew Ho, Harvard University; Morgan Polikoff, University of Southern California

Assessment for Admission, Placement, and Outcomes in Higher Education
Wayne Camara, Law school admissions council; Ou Lydia Liu, ETS; Krista Mattern, National Board of Medical Examiners

Assessment for Licensing and Certification
Melissa Margolis, NBME; Rebecca Lipner; Chad Buckendahl, ACS Ventures, LLC

Discussants:
Mark Wilson, University of California, Berkeley
Emily J. Shaw, College Board

178. Electronic Board Session: AI

Electronic Board Session
1:30 PM - 3:00 PM
Hilton Denver City Center Lower Level 2: Colorado GH

Presentations:

1. **Designing a pipeline for item development using NLP, LLMs and Gen-AI**
   Tzur Karelitz, National Institute for Testing & Evaluation (NITE); Effi Levi, National Institute for Testing & Evaluation (NITE); Joni Sarig, National Institute for Testing & Evaluation (NITE); Yael Shor, National Institute for Testing & Evaluation (NITE); Roi Shillo, National Institute for Testing & Evaluation (NITE)
   We present a pipeline for creating Grammar-in-Context items for a high-stakes English test. We divided the item development process into multiple separate tasks and identified the appropriate tools for each one. This was accomplished by utilizing NLP tools (e.g., constituency parser), LLM-based tools (e.g., BERT), and Gen-AI models (e.g., chatGPT).

2. **Using Large Language Models for Automatic Item Generation: Validation for TIMSS Grade 4**
   Marek Muszyński, Institute of Philosophy and Sociology of the Polish Academy of Sciences; Artur Pokropek, IFiS PAN; Hubert Plisiecki, Institute of Philosophy and Sociology of the Polish Academy of Sciences; Tomasz Żółtak, IFiS PAN
   This project aims to utilise large language models (LLMs) for the automatic generation of items for the Trends in International Mathematics and Science Study (TIMSS) Grade 4 assessment. The automatically generated items will be compared empirically to original TIMSS items with respect to their psychometric properties and expert evaluations.

3. **Automated Item Generation for Reading Assessments**
   Ummugul Bezirhan, Boston College; Erin Wry, Boston College; Matthias von Davier, Boston College
   This study explores automated item generation (AIG) in PIRLS framework, utilizing GPT models to automate the creation of reading comprehension items. By integrating AI-driven methods with PIRLS item writing guidelines, this research seeks to demonstrate an efficient and streamlined approach to item development to reduce the burden on experts.

4. **Exploring Human and Automated Scoring System Agreement for Short Answer Questions**
   Stephanie Iaccarino, NBME; Brian Clauser; Polina Harik, NBME; Michael Kane; Peter Baldwin, National Board of Medical Examiners; Yiyun Zhou
   When evaluating and finetuning automated scoring systems, treating expert scores as true overlooks and obscures variation across experts. In this study, free-text responses were independently scored by two groups of experts and used to twice finetune a scoring system. Strong within-group expert/system agreement and modest between-group expert/expert agreement were observed.

5. Scoring Math Constructed Response Prompts with a Hybrid Automated Scoring Model
Lisa Eggers-Robertson
This study evaluates the use of a combination of rules-based and text conversion techniques as a valid automated scoring method for responses to math prompts. We compare the scoring engine's performance against human scores in terms of the reliability and potential bias of using this type of automated scoring.

6. Optimizing Large Language Models for Automated Scoring of Constructed Response Mathematics Items
Elana Margaret Perkoff, University of Colorado Boulder; Ruhan Circi, American Institutes for Research
This study explores the performance of lightweight and complex Large Language Models (LLMs) in scoring constructed response mathematics items, emphasizing training sample size, response score categories, and item difficulty. It contributes to the growing literature and provides insights into key factors for effective operational use of LLMs in automated scoring.

7. Evaluating Large Language Models in Automated Online Comment Review
Ting Wang; Ying Du, American Board of Pediatrics; Karen Hoeve, American Board of Pediatrics
This study compared Large Language Models' (LLMs: GPT-4o) performance against traditional Machine Learning (ML) methods in detecting problematic assessment questions from online comments. Using F1 score to address class imbalance, GPT-4o (F1 score = 0.93) outperformed all typical traditional ML methods considered (F1 scores ranges 0.62 to 0.85).

8. Enhancing Systematic Reviews with Large Language Models: Using GPT-4 and Kimi
Danielle (Dandan) Kaptur, Pearson; Yue Huang, Measurement Incorporated; Xuejun Ji; Yanhui Guo, University of Illinois Springfield; Bradley Kaptur, HSHS Saint John's Hospital
This research delved into GPT-4 and Kimi, two Large Language Models (LLMs), for systematic reviews. We evaluated their performance by comparing LLM-generated codes with human-generated codes from a peer-reviewed systematic review on assessment. Our findings suggested that LLMs' performance fluctuates by data volume and question complexity for systematic reviews.

9. The Utilization of Algorithmic Mathematics Models in Automated Scoring for Diagnostic Feedback
Jennifer Novak, Pearson; Lisa Eggers-Robertson
This study explores how an automated scoring tool can efficiently provide diagnostic feedback using algorithmic mathematics models in automated scoring. Our results show that the automated scoring tool can report diagnostic feedback along with a test-taker's score.

10. Improving Zero-Shot Classification of Alerting Content Essays with Automatically Generated Descriptions
Sungjin Nam, ACT, Inc
We present a method for automatically detecting alerting papers in standardized essay tests without labeled data. Preliminary results show that the zero-shot learning approach outperforms supervised learning models trained with out-of-domain datasets. Descriptions for alerting content categories generated by a large language model provide better classification performance than human-authored descriptions.

11. Using Rasch Modeling to Validate AI Scoring Models
Steven McGee, The Learning Partnership; Yaurie Hwang, The Learning Partnership; Willow Kelleigh, The Learning Partnership
This study explores the development and validation of AI scoring models for assessments related to the Exploring Computer Science curriculum. Using BERT, AI models were trained to replicate human scorers and validated using Rasch modeling. Results showed the AI models approximated human scorers, offering opportunities for timely, formative feedback.

12. An Evaluation of Automated Item Scoring in the Context of Inter-Rater Reliability
Kate Nolan, Amira Learning; Youngsoon Kang, Amira+Istation; Ran Liu, Amira Learning
This preliminary study evaluates the scoring reliability of a machine learning (ML) model that assesses student speech, comparing it to human raters. Results indicate strong human-machine correlation, suggesting the ML model could be a viable alternative for consistent and efficient scoring of classroom reading assessments.

13. A Mixed-Methods Study on the Application of ChatGPT in AP Chinese Teaching
Dan Song, The University of Iowa; Frederick Poole, Michigan State University
This study explores the adoption of ChatGPT by AP Chinese teachers for course preparation within the Technological Pedagogical Content Knowledge (TPACK) framework. It assesses the impact on technological knowledge (TK), pedagogical knowledge (PK), and content knowledge (CK) through a mixed-method approach, focusing on integrating AI into high-stakes language instruction.

14. Machine Learning-Based Survival Categorization of Response Time
Ye Lin, Ascend Learning; Kari Hodge, Ascend Learning
In this study, we leveraged machine learning to predict response time for nursing exam items with different scoring methods. By analyzing linguistic features and word embeddings from item stems and options, we compared the performance of various algorithms. Our findings offer valuable insights for pretesting and test construction practices.

15. Using natural language processing to identify enemy items
Aijun Wang; Yu Zhang, The Federation of State Boards of Physical Therapy
Manual identification of enemy items during the form review process is both time consuming and prone to inconsistencies. This study demonstrated the effectiveness of NLP techniques in identifying enemy items to help with the form review process.

16. Quantifying response quality of RAG chatbot on NAEP Assessment Information
Ting Zhang, American Institutes for Research; Luke Patterson, American Institutes for Research; Magdalen Beiting-Parrish, Federation of American Scientists; Blue Webb, American Institutes for Research; Bhashithe Abeysinghe, American Institutes for Research; Paul Bailey, American Institutes For Research; Emmanuel Sikali, NCES
This study evaluates the Ask NAEP, an information retrieval chatbot enhanced with Retrieval-Augmented Generation (RAG) to improve response quality for user queries on NAEP information. Experiments show that GPT-4o outperformed GPT-3.5. Incorporating retrieval into the pipeline further enhanced performance. The RAG approach significantly enhances chatbot accuracy and user experience.

179. Admission, Placement, and Outcomes in Higher Education

Individual Paper Session
1:30 PM - 3:00 PM
Hilton Denver City Center Lower Level 2: Denver 5-6

Chair:
Roza Nalbandyan

Presentations:
Evaluating Articulation Agreements: Metrics for Improving Credit Transfer and Student Success
Sarah Goulart, UC Berekeley; Zachary Pardos, University of California, Berkeley
This paper introduces metrics—Articulation Coverage Score (ACS), the AI-enhanced ACS, Extra Credit Granted, and Unused Articulation—using State University of New York (SUNY) data. We examine 2- and 4-year institutions' coverage, credit granting, and highlight how these metrics identify areas for improvement to reduce credit loss and enhance transfer experiences.

Subgroup Performance Changes Pre- and Post-Covid: SAT Scores, HSGPA, and FYGPA
Paul Westrick, College Board; Jessica Marini, College Board; Emily Shaw, College Board
Previous research has examined undergraduate student performance changes pre- and post-covid, overall and by institutional subgroupings. Standardized test scores tended to decline while high school and undergraduate grades increased. The current study expands upon this research by examining changes across student subgroups of interest, notably underrepresented minorities, and first-generation students.

Standardized Testing and Social Equity: Assessing Recent Changes in Chile's University Admissions
David Torres Irribarra, Pontificia Universidad Catolica de Chile; Maria Santelices
We examine the impact of changes to Chile's national standardized university admissions tests on fairness. We analyze two years before the changes (2018/2019) and two after (2023/2024). We examine gaps between genders, public/private schools, and vocational/academic schools, and fairness using DIF methods. Results inform the change's effectiveness in enhancing fairness.

Predictive Validity Changes in HSGPA and ACT Scores after the Pandemic Onset
Edgar Sanchez
This study evaluates the predictive validity of high school GPA and ACT scores on first-year college GPA before and after the onset of COVID-19. Results show a significant change in HSGPA's predictive power post-2020, while ACT scores remained stable. The study highlights the compensatory nature of using both measures together.

Discussant:
Lissette Tolentino, University of Florida

180. Teacher Education

Individual Paper Session
1:30 PM - 3:00 PM
Hilton Denver City Center Lower Level 2: Denver 3

Chair:
Tara Kintz, Michigan Assessment Consortium

Presentations:
Enhancing Teaching and Learning: A Mixed Methods Study on 21st Century Skills
Merve Şahin Kürşad, TED University; Aybuke Durmus, TED University; Ilgun Dibek Munevver, TED University
This study uses a convergent mixed methods design to assess 21st-century skills training for pre-service teachers. By integrating quantitative and qualitative data, it highlights significant improvements in teaching practices and learning skills, providing a comprehensive evaluation of the training's effectiveness and impact on participants.

Communicating Effect Sizes to Teachers: A Replication in the U.S.
Sanford Student, University of Delaware; Benjamin Shear, University of Colorado Boulder; Wyatt Read, University of Delaware
Recent work investigated the clarity and perceived effectiveness of different effect size translations among U.K. teachers: percentile gain, months of progress, percent proficient, Cohen's U3, and raw mean difference in scores. We present a U.S. replication, providing new evidence on how effect sizes can be communicated to practitioners most effectively.

Relationship between Early Childhood Montessori Teacher Math Beliefs & Confidence and Classroom Environment
Angela Murray, University of Kansas, Carolyn Daoust, University of Kansas; Heather Gerker, University of Kansas; Amelia Murray, University of Kansas
This paper summarizes one component from a larger Montessori fidelity pilot study. Participating

teachers completed a math beliefs and confidence survey to investigate the relationship of teachers' math affinity with classroom math practices to uncover opportunities to enhance Montessori math learning.

Exploring Objective Measures to Detect Math Teacher Bias Using Multivariate Assessment Data
Tamlyn Lahoud, University of Georgia; Constanza Mardones-Segovia, University of Georgia; Nohwon Park, University of Georgia; Shiyu Wang, University of Georgia; Allan Cohen, University of Georgia; Yasemin Copur-Gencturk, University of Southern California
This study analyzes multivariate data on teachers' ratings of students' math performance and ability, using multiple statistical methods to reveal differences in teacher evaluations based on student gender or race. Findings show that teacher bias can be effectively monitored through objective measures, enhancing test fairness without relying on subjective evaluation.

Discussant:
Courtney Bell

181. Validity, Validation, and Fairness Frameworks and Considerations

Individual Paper Session
1:30 PM - 3:00 PM
Hilton Denver City Center Lower Level 2: Colorado IJ

Chair:
Jungwon Kyung, University of Massachusetts, Amherst

Presentations:
Content and Cognition Traits of High Quality Valid Test Items
Marjorie Wine, Accessible Teaching, Learning and Assessment Systems (ATLAS), University of Kansas; Christy Glore, ATLAS, University of Kansas; Alexander Hoffman, AleDev Research & Consulting
This paper offers a set of principles—beyond item hygiene issues—to apply in item development and item evaluation during the test development process that focuses on the substantive content of items and their ability to elicit evidence of the targeted cognition for the range of typical test takers.

Validity Under Scrutiny: Review of Educational and Psychological Textbooks Post-1999 Standards
Susan Embretson, Georgia Institute of Technology; Haiying Long, University of Kansas
This study examines the concept of validity in 110 educational and psychological textbooks published between 2001 and 2024 for inclusion of changes in the Standards for Educational and Psychological Testing since 1999. It found that only a small number of textbooks adopt the contemporary view of validity.

Using Think Aloud, Surveys, and Interviews to Validate Cognitive Processes for an Assessment
Jay Thomas, ACT Education Corp; Scott Johanningmeier, ACT Education Corp; Stacy Dreyer, ACT Education Corp; Chuqing Zhang, Mt. Holyoke
This paper explores how to gather validity evidence for cognitive processes that support claims of what test takers know and can do by using cognitive labs including think-alouds and interviews. We show how coding differences in the use of targeted and maladaptive skills lead to different performance outcomes.

Fairness in Quantitative Measures used in Educational Effectiveness Research Studies
Corinne Huggins-Manley; Madeline Browy, University of Florida; Raegan DiRenzo, University of Florida; Grey Chapin, University of Florida; Amanda Siu, University of Florida; Balaji Iyer, University of Florida
This study explores the extent to which researchers provide evidence for fairness in the measurement scores used to evaluate effectiveness in educational research publications. We answer the research question: What is the frequency of What Works Clearinghouse (WWC) studies on educational effectiveness that evaluate the fairness in their outcome measurements?

Discussant:
Thanos Patelis, University of Kansas

182. Interrogating the Value of Classroom-external Assessment: A Foundational Discussion for the Field

Organized Discussion
1:30 PM - 3:00 PM
Hilton Denver City Center Lower Level 1: Penrose 1

This collaborative, interactive, organized discussion presents three perennial questions about external assessments, including large-scale state tests: 1) Who benefits from external assessment in the current landscape, and under what conditions? 2) How might we build on what is known about how external tests can promote learning while also serving the purposes of measurement? 3) How can external assessment results and reports be used to inform instruction, student learning, and families' support of students? The session will bring a diverse set of partners to the table to share information across partner groups, increase understanding, question assumptions, and generate new ideas. Presenters will provide a foundational overview and background information on each question. Session attendees will play an active role in sharing their expertise and wonderings in small group discussions. The results of the session will be used to spur additional collaborations among different groups within and beyond NCME.

Chair:
Dustin Van Orman, Western Washington University

Moderator:
Dustin Van Orman, Western Washington University

Presenters:
Yi-Chen Wu, National Center on Educational Outcomes, University of Minnesota
Shenghai Dai, Washington State University
Heidi Andrade, University of Albany
Dustin Van Orman, Western Washington University

183. Addressing Items with Multiple-DIF Flags and Other Fairness Issues across Assessment Sectors

Organized Discussion
1:30 PM - 3:00 PM
Hilton Denver City Center Lower Level 1: Penrose 2

High-stakes testing programs routinely evaluate items for DIF. Item DIF can be positive (advantaging a focal group), negative (advantaging a reference group), or mixed (i.e., nonuniform DIF, where low- or high-performing students from either the focal or reference group is advantaged, while the other group is disadvantaged). How test developers analyze DIF and use its results can vary across programs. However, one ubiquitous aspect of DIF is that it is typically analyzed over multiple focal-reference (FR) groups (e.g., Female-Male, Black-White). During these analyses, a nontrivial number of items can be flagged multiple times for having significant DIF over the different FR group pairs. Too often ignored in practice is the question of whether practitioners should conduct special analyses or have unique review, retention, and/or retest policies for these items. The five panelists for this session will consist of leading representatives across different assessment spaces (K-12, Admissions, Credentialing, Academics, etc.) who will discuss the special implications associated with multiple-DIF items. In addition to discussing the analysis and policy issues, panelists will also share frequency data and exemplar items as well as discuss the implications for item development, test assembly, and the validity of test-score inferences and interpretations.

Chair:
Norman Bishop, LSAC

Moderator:
Anna Topczewski, Law School Admissions Council

Presenters:
Marc Julian, DRC
Susan Davis-Becker, ACS Ventures, LLC
Michael Walker, HumRRO
Ye Tong, National Board of Medical Examiners
Kyndra Middleton, Howard University

184. Applications of Large Language Models to Automated Scoring and Feedback
Coordinated Paper Session
3:15 PM - 4:15 PM
Hilton Denver City Center Lower Level 2: Colorado IJ

The NCME Annual Meeting focuses on how measurement can benefit society. As impressive as the capabilities of Large Language Models (LLMs) are, attempts to leverage LLMs to benefit students and educators have faced challenges. We know that stakeholders in education are inherently cautious, so any applications of LLMs must be subjected to a rigorous validation agenda before consideration in any large-scale assessment program. Directly prompting LLMs to obtain scores and feedback has so far failed to meet our standards. This session is a showcase of innovative alternative ways that LLMs can provide value to the scoring and feedback pipeline, either by providing useful features or the use of fine-tuned models. These higher-level features, annotations, and fine-tuning are shown improve pipelines, make them more robust, and provide useful information to students and educators a manner that admits careful validation. This session includes applications of proprietary generative language models, such as GPT-4, fine-tuned open-source models, such as Llama-3, and traditional encoder-decoder models like BERT and XLNet. This includes research from Cambium Assessment, Performance Assessment Analytics, and College Board.

Session Organizer:
Christopher Ormerod,

Presentations:
Automatic Detection of Invalid Essay Response Categories Using ASAP Data and GPT-4o
Ryan Lerch; YoungKoung Kim, The College Board

Automated Text Scoring in the Age of Generative AI for the GPU-poor
Christopher Ormerod, Alexander Kwako, Cambium Assessment

Extending Human and AI Annotations beyond Argumentative Writing
Amy Burkhardt, Cambium Assessment; Sherri Woolf, Consultant; Mackenzie Young, Christopher Ormerod, Susan Lottridge, Cambium Assessment; Gitit Kehat, Cambium Assessment

Discussant:
Okan Bulut, University of Alberta

185. Development and Implementation of RTD Task Models for Large Scale Assessment Content Development
Coordinated Paper Session
3:15 PM - 4:15 PM
Hilton Denver City Center Lower Level 2: Denver 3

Task models are the final structure that Evidence Centered Design (ECD) offers before the black box of content development, but the literature says remarkably little about them, or instrumentalization more broadly. This symposium explores the evolution of one approach to task models. The first paper addresses the historical development and reasoning that has brought RTD Task Models to their current form. This includes their design, the particular problems of content development that they respond to and their goals for improving the process. The second paper dives deeply into the collaborative process of unpacking new NGSS-based alt standards for use as the targeted cognition (assessment targets)—the foundational first step in writing RTD Task Models. The third paper examines how developing and using RTD Task models influenced the relationships and information flows between client and test development vendor, among content development teams and for item writers. This resulted in greater efficiencies in acceptance rates of developed items. Taken together, these papers demonstrate how task models can leverage content development professionals (CDPs) as engines of organizational learning and developing evidence of validity for their assessments.

Session Organizer:
Alexander Hoffman, AleDev Research & Consulting

Chair:
Marjorie Wine, Accessible Teaching, Learning and Assessment Systems (ATLAS), University of Kansas

Presentations:
RTD Task Models: Principles, Originating Problems, Purposes and Design
Marjorie Wine, Accessible Teaching, Learning and Assessment Systems (ATLAS), University of Kansas; Alexander Hoffman, AleDev Research & Consulting

Recognizing Targeted Cognition: Collaborative Practice in Unpacking NGSS-based Alternative Science Standards
Christy Glore, ATLAS, University of Kansas; Cara Olson-Rickel, ATLAS, University of Kansas; Caroline Sloan, ATLAS, University of Kansas; Shari Tolan, ATLAS, University of Kansas

RTD Task Model Conjectures: Initial Lessons from a 5th Grade Mathematics Assessment
David Whitcomb, ATLAS, University of Kansas; Alexander Hoffman, AleDev Research & Consulting

Discussant:
Ellen Forte, edCount, LLC

186. Within-year Growth Measures and Learning Inferences
Coordinated Paper Session
3:15 PM - 4:15 PM
Hilton Denver City Center Lower Level 2: Denver 4

In contrast to single test events, repeated measures can help educators understand past performance, predict future performance, and in some cases inform instructional decisions—all potential benefits to educators and society at large. Increasingly testing programs are providing repeated testing opportunities for learners, supporting new formulations of achievement and growth inferences. For instance, cumulative scoring approaches combine scores from multiple within-year test events providing formulations of achievement that are related to, yet distinct from end-of-year summative scores. Also, repeated within-year measures produce growth estimates distinct from commonly implemented across-year, spring-to-spring growth estimates. Presentations will cover three topics

relevant to testing programs using within-year repeated measures. The first study compares the precision and measurement sensitivity of simple gain scores, student growth percentiles, and multilevel growth models. The second study compares the ability of four IRT-based growth models to recover latent growth variables. The third study will apply an IRT-based model, the individualized-student slope Additive Factors Model, to estimate task difficulties, student slopes, and topic-specific slopes using data from a longitudinal adaptive learning system. Collectively these studies present different modeling approaches and different uses of repeated measures, each with their own theory of action on how the measures will benefit users.

Session Organizer:
Garron Gianopulos, Cambium Assessment

Chair:
Garron Gianopulos, Cambium Assessment

Presentations:
The Precision and Sensitivity of Three Within-year Growth Metrics
Garron Gianopulos, Cambium Assessment

A Comparison Between Longitudinal IRT Models for Combined Overall Scoring
Maryam Pezeshki, Cambium Assessment Inc.; Shuqin Tao, Cambium Assessment Inc.; Stephan Ahadi, Cambium Assessment Inc.

Eliciting Assessment From Adaptive Learning Tools
mary Ann Simpson, Carnegie Learning; Steven Fancsali, Carnegie Learning; Steven Ritter, Carneg

Discussants:
Chun Wang, University of Washington
Nathan Dadey, The National Center for the Improvement of Educational Assessment

187. Advances in Item Calibration for CAT

Individual Paper Session
3:15 PM - 4:15 PM
Hilton Denver City Center Lower Level 2: Colorado CD

Chair:
Jaime Malatesta, GMAC

Presentations:
CD-CAT Online Item Calibration with Practical Constraints
Jing Huang; Yuxiao Zhang, Purdue University ; Jason Morphew, Purdue University; Jayson Nissen, LASSO; Ben Van Dusen, Iowa State University; Hua-Hua Chang, Purdue University
This study aims to investigate the effectiveness of incorporating practical constraints into the CD-CAT online calibration procedure. Preliminary results indicate that adding content area constraints, increasing sample size and using smaller item parameters improve item pool utilization while maintaining accurate new item estimates. This research helps teachers provide targeted instruction.

Termination Criteria for Online Calibration of Polytomous Item Sets
Zhuoran Wang, NCSBN; William Muntean, National Council of State Boards of Nursing; Joe Betts, NCSBN
In Computerized Adaptive Testing (CAT), online calibration improves pretesting efficiency, yet the full potential of adaptive principles is not fully utilized, especially for terminating pretests. This study evaluates three pretest termination criteria for CATs featuring polytomous items and item sets, offering a comprehensive discussion of their practical applicability.

The Impact of Item Calibration Accuracy on Computer Adaptive Testing
Jie Li
This paper studies how item calibration accuracy impacts Computer Adaptive Testing (CAT). Traditional thoughts are the more accurate item calibration to recover item parameters, the better for operational test scoring. This study shows that it may not always be true for computer adaptive testing.

Discussant:
Jonathon Weeks, American Institutes for Research

188. Classroom and Formative Assessment

Individual Paper Session
3:15 PM - 4:15 PM
Hilton Denver City Center Lower Level 1: Penrose 1

Chair:
Roti Chakraborty, Georgia State University, American Institutes for Research

Presentations:
Contextualising Formative Assessment: A Teacher-Researcher Collaboration in Tanzanian Mathematics Classrooms
Therese Hopfenbeck, University of Melbourne; Joshua McGrane, University of Melbourne; Anjum Halai, Aga Khan University
This study explores the implementation of formative assessment in Tanzanian primary schools. Through collaborative workshops and classroom observations, researchers and teachers adapted existing self-regulated learning and assessment models to fit large class sizes. Findings highlight the importance of contextual knowledge in implementing formative assessment practices and the value of collaborative research-partnerships.

Framing Assessment Reform for the Caribbean using Decolonization Theories
Jerome De Lisle, University of the West Indies - St. Augustine
This paper explicates decolonization theories to frame student assessment reform in the Caribbean, an archipelago of postcolonial nation states. The constructed framework compares well to Euro-American change theory but better explains current system architecture and provides a more focused pathway for achieving greater equity, with liberation and justice.

Rapid Guessing Methodology in an Operational Formative Assessment Context Leveraging Diagnostic Measurement
Ruchi Sachdeva, Pearson; Laine Bradshaw, Pearson
This study examines rapid guessing in Navvy, a K-12 formative assessment system using diagnostic measurement (DM) models. Rapid guessing threatens assessment reliability and validity by distorting student competency diagnoses. We evaluate response time thresholds to identify disengagement while minimizing misclassification, offering solutions to improve the accuracy of formative assessments.

Discussant:
Sandra Liliana Camargo Salamanca, University of Illinois Urbana-Champaign

189. Addressing Equity-Related Educational Challenges

Individual Paper Session
3:15 PM - 4:15 PM
Hilton Denver City Center Lower Level 1: Penrose 2

Chair:
Omaya Horton, University of Massachusetts Amherst

Presentations:

School Characteristics and Socio-Economic Status in States with High- vs Low-Performing Students
Eric Asare, Old Dominion University; Daniel Edi, NC Department of Public Instruction
This study examines the role of socio-economic status and school characteristics in states with high- vs. low-performing students, based on NAEP assessment data on students in grades 4 and 8 for math and reading. This research provides insights that can inform policy decisions aimed at improving educational outcomes and equity.

Segregation in Mathematics Classrooms: Examining Indicators of Heterogeneity
Roza Nalbandyan
English Learners are more likely to succeed academically in linguistically heterogeneous classes. We develop classroom linguistic heterogeneity indicators on 8.990 students' data using various statistical approaches and examine their limitations and possibilities. All indicators have advantages and disadvantages. They can help schools and policymakers develop more effective assignment practices.

Going Beyond the Score: How Assessments Can Address Key Educational Challenges
Evelyn Johnson, Riverside Insights; Emily Parker; Jennifer Robitaille, Riverside Insights
This session presents a series of studies examining practical applications of educational assessments, used in partnership with school districts to address chronic absenteeism, support English learners, and to support the whole-child. The goal is to highlight how assessments can be used to help address critical challenges in schools.

Discussant:
Michael Rodriguez, University of Minnesota

190. Licensure and Certification

Individual Paper Session
3:15 PM - 4:15 PM
Hilton Denver City Center Lower Level 2: Colorado AB

Chair:
Kelley Stethen, ACS Ventures

Presentations:

Recommender Systems for Personalized Item Selection in Assessments of Medical Knowledge
Kelly Edwards, American Board of Pediatrics; Andrew Dwyer, American Board of Pediatrics
This study introduces a novel recommender system-based algorithm for item selection in longitudinal assessments for maintenance of certification. Utilizing collaborative filtering and a modified weighted penalty model, it tailors exam content to individuals based on predicted relevance to clinical practice. Results demonstrate the approach optimizes practitioner-rated relevance of selected items.

Predicting Success: Joint Modeling of Retake Behaviors in Teacher Licensure Assessment
Caitlin Tenison, ETS; Teresa Ober, ETS; Geoffrey Phelps, ETS
We present joint longitudinal and time-to-event models to capture the various factors that predict the distance of test takers from passing across retakes, while accounting for the effect of right-censoring due to individuals who cease retaking. We observe that poor performance and worsening scores significantly reduces the hazard of passing.

Praxis to Inform Practice: Generating Insights on Topic Difficulty to Support Teachers
Teresa Ober, ETS; Geoffrey Phelps, ETS; Caitlin Tenison, ETS
We investigate topic category performance on the Praxis Elementary Multisubject exams. Study 1 used classification trees and identified Science and Social Studies as challenging areas. Study 2 used latent class analysis to identify examinee performance profiles. Both studies offer insights for targeted interventions to strengthen teacher competencies.

Discussant:
Susan Davis-Becker, ACS Ventures, LLC

191. Reliability, Precision, and Generalizability

Individual Paper Session
3:15 PM - 4:15 PM
Hilton Denver City Center Lower Level 2: Colorado GH

Chair:
Mingfeng Xue, University of California, Berkeley

Presentations:

Item reduction for digital literacy assessment: Perspectives from content-expert, psychometrics, and machine-learning
He Ren, University of Washington; Qianqian Pan, National Institute of Education, Nanyang Technological University; Yuxiao Zhang, Purdue University; Weicong Lyu, University of Macau
Digital literacy (DL) is a crucial skill in the modern world. This study streamlined a DL assessment using data-driven methods grounded in psychometrics and machine learning. By comparing the data-driven shortened DL assessment with the expert-developed insight-driven version, we offered diverse insights into DL evaluation and item reduction methodologies.

Exploring Selection of Reference Indicators for LST Methods Minus One Models
Tingting Chen, The University of Iowa; Walter Vispoel, The University of Iowa
We compared methods minus one (M – 1) latent state-trait models using different reference indicators based on exploratory factor analysis loadings. Results revealed that choosing the highest-loading item provided an effective strategy for selecting the best reference indicator in the absence of theoretical considerations or other empirical data.

When Are Factor Scores or Sum Scores More Reliable?
Winnie Wing-Yee Tse, Vretta Inc.; Hok Chio Lai, University of Southern California
While factor scores theoretically are at least as reliable as sum scores, their reliability is affected by sampling errors in scoring weights. In this presentation, we derive the finite-sample reliability of factor scores and introduce two procedures for determining when factor scores are more reliable than sum scores.

Discussant:
Jessalyn Smith, DRC

192. Standard Setting

Individual Paper Session
3:15 PM - 4:15 PM
Hilton Denver City Center Lower Level 2: Denver 5-6

Chair:
Serife Bilgic-Erdem, William & Mary

Presentations:

A Simulated Comparison of Embedded Standard Setting methods on Cut Score Estimation
Sangdon Lim, Cambium Assessment; Christina Schneider
Embedded Standard Setting methods use rater data containing achievement level descriptor (ALD) assignments to calculate ALD cut scores. Raters can vary in expertise and style in assigning ALDs to items, which can affect cut score estimates. Count and weighted Embedded Standard Setting were compared across rater patterns for estimation performance.

A Multivariate Generalizability Theory Approach to Assertion-Mapping Procedure in Standard Setting
Yi-Fang Wu, Cambium Assessment, Inc.
We utilize multivariate generalizability theory to evaluate mapping consistency in the assertion-mapping procedure for setting performance standards in cluster-based assessments. We conduct single- and multi-facet G studies to disentangle variance components, derive a mapping consistency index, and evaluate AMP across elementary, middle, and high school panels for various states.

Validating proficiency levels set via Bookmark standard setting using Latent Class Analysis
Kari Hodge, Ascend Learning; Ye Lin, Ascend Learning; Nicole Zapparrata, Ascend Learning
The Bookmark method requires SMEs to make judgments about exam items. Given the subjectivity of human rating, it's essential to validate cut scores. The current study employed LCA to investigate the validity of the Bookmark method. The LCA 2-class solution and Bookmark methods showed a substantial level of agreement.

Discussant:
Daniel Lewis, Creative Measurement Solutions

INFORMAL PRESIDENTIAL CLOSING RECEPTION

Saturday, April 26

4:30 PM TO 6:00 PM

Location to be announced at the Presidential Address on Friday morning

PARTICIPANT INDEX

(Last Name, First Name, Session Number)

Ketan, 061, 103
Abeysinghe, Bhashithe, 178
Abroampah, Doris, 114
Abulela, Mohammed, 075, 103
Acharya Julien, Nisha, 117
Ackerman, Terry, 045, 089
Adegoke, Hope, 168
Ahadi, Stephan, 042, 095, 186
Aiello, Liam, 150
Akande, Christiana, 032, 146
Akbay, Lokman, 131
Alahmadi, Sarah, 054
Alarcón, Eduardo, 144
Albano, Tony, 029, 038, 073, 122, 149,176
Alila, Nancy, 131
Allen, Jeff, 037
Almond, Russell, 003, 063
Amoateng, Ernest, 160
An, Lily, 065, 120, 149
Anderson, Elizabeth, 031
Andrade, Heidi, 050, 067, 096, 182
Andrade-Lotero, Luis Alejandro, 040
Andrews-Todd, Jessica, 026, 064
Anghel, Ella, 124
Anguiano-Carrasco, Cristina, 091, 105
Araneda, Sergio, 060, 088, 129, 130
Arboleda, Juan, 168
Arias, Angel, 070
Armoush, Nour, 148
Arslan, Burcu, 133, 152
Arthur, Ann, 037
Arthur, David, 122
Asamoah, Nana Amma, 073
Asare, Eric, 189
Athanases, Steven
Attali, Yigal, 070, 079
Au, Wayne, 055
Augustsson, Keith, 157
Avila, Natalia
Avraam, Demetris, 133
Ayers-Wright, Elizabeth, 146
Badrinarayan, Aneesha, 117, 163
Bahry, Louise
Bailey, Alison, 168
Bailey, Jennifer, 124, 133
Bailey, Paul, 001, 115, 178
Baker, Doris Luft, 033
Baker, Eva, 039
Baldwin, Peter, 093, 178
Balisciano, Nicholas, 051
Bamat, David, 049
Banjanovic, Erin, 042, 095, 167
Bao, Yu, 073, 103, 114, 168
Baral, Kushmakar, 031, 122
Barber, Justin, 090, 101
Barnes, Sophie, 120
Barnum, Stuart, 106, 114
Baron, Zach, 114
Barry, Carol, 093
Bates, Shaun, 042
Bay, Luz, 087, 105
Beach, Sarah, 098
Beard, Jonathan, 052
Bechard, Sue, 154
Becker, Kirk, 056
Becker, Lee, 040, 090
Bediwy, Ahmed, 103, 136, 156
Behan, Bradley, 090
Beiting-Parrish, Magdalen, 027, 098, 150, 164, 178
Bell, Courtney, 097, 180
Bellows, Martha, 101
Belov, Dmitry, 159
Belur, Vinetha, 102
Belzak, William, 056, 070, 134
Bennett, Randy, 044, 050, 163
Bergner, Yoav, 147
Bernacki, Matthew
Bernard, Steve, 131
Beswick, Bruce, 064
Betts, Joe, 093, 139, 187
Beuoy, Aaron, 051
Bezirhan, Ummugul, 074, 178
Bhadra, Dhiman
Bhandari, Shreya, 111
Biancarosa, Gina, 061, 176
Bilgic-Erdem, Serife, 099, 192
Binzak, John, 083
Bishop, Kyoungwon, 083, 129, 157
Bishop, Norman, 183
Bohrnstedt, George, 049
Bolender, Brad, 064
Bolt, Daniel, 037, 047, 089, 103
Bonge, Nicole, 114
Bonifay, Wes, 007
Bonk, Bill, 134
Bonner, Sarah, 050
Born, Marise, 031
Botter, Preston, 103, 168
Boughton, Keith, 051, 174
Bovaird, James, 103, 158
Boyd, Aimee, 176
Boyd, Brittany
Boykin, Allison, 081, 123, 147
Boykin, Malik, 044
Bradley, Kelly, 031
Bradshaw, Laine, 161, 188
Braun, Henry, 132
Brennan, Robert, 130
Brettschneider, Alli, 039
Briggs, Derek, 045, 113, 144, 167
Broer, Markus, 153
Brookhart, Susan, 050, 177
Brown, Emily, 145
Brown, Julianna, 107
Brown, Rachel, 096
Brown, Richard, 083, 145
Brown, Terran
Browy, Madeline, 181
Brückner, Sebastian, 132
Brunetti, Matthew, 176
Brunskill, Emma, 111
Bryer, Jason, 040, 096
Buckendahl, Chad, 068, 096, 138, 151, 177
Buckley, Jack, 064
Bueso, Leah, 039
Bufferfuss, Reese, 135
Bulut, Okan, 027, 048, 148, 152, 184
Burkhardt, Amy, 036, 074, 090, 123, 184
Burstein, Jill, 142
Buzick, Heather, 025
Cai, Li, 033, 044
Cai, Liuhan, 099, 174
Cai, Yan, 054, 104, 136
Cain, Jennifer, 102
Calderón, John Alexander, 032
Caliço, Tiago, 001, 092
Camara, Wayne, 177
Camargo Salamanca, Sandra Liliana, 061, 188
Campbell, Ian, 073
Can Aran, Ozge
Cancado, Luciana, 095
Cao, Yichong, 087, 126
Cardwell, Ramsey, 134
Carpenter, Blaire, 063
Carrasco, Diego, 103
Casabianca, Jodi, 005, 101, 164
Castillo, Carolina, 146

PARTICIPANT INDEX

(Last Name, First Name, Session Number)

Cayumán, Carlos, 144
Celic, Vuk, 114
Cervantes, Víctor H, 032
Chajewski, Michael, 095
Chakraborty, Roti, 150, 188
Chalhoub-Deville, Micheline, 134
Cham, Heining, 061
Chang, Chi, 093
Chang, Hua-Hua, 187
Chang, Joshua, 073
Chang, Wanchen, 081
Chao, Hsiu-Yi
Chapin, Grey, 181
Chavez, Carlos, 098
Chawla, Kamal, 051
Chen, Fangbin, 054
Chen, Binhui, 051
Chen, Chi-Chen, 031
Chen, Hong, 146
Chen, Hui-Fang, 083
Chen, Jianshen, 116
Chen, Jing, 029
Chen, Jinsong, 168
Chen, Jyun-Hong
Chen, Qipeng, 004, 059
Chen, Siyuan, 032
Chen, Tingting, 191
Chen, Xiuhan, 146
Chen, Ying, 157
Chen, Yinghan, 112
Chen, Yunxiao, 053
Chen, Zhijun, 103
Cheng, Britte, 098
Cheng, Ying, 059, 104, 119, 136
Chiu, Chia-Yi, 157
Cho, Youngmi, 157
Cho, YoungWoo, 037, 076
Choe, Edison, 029, 052, 073
Choi, Hye-Jeong, 135
Choi, Ikkyu, 111, 152
Choi, Jinah, 033, 100
Choi, Jinnie, 122
Choi, Mihye, 167
Choi, Seung, 148
Christensen, Laurene, 154
Christensen, Wendy, 125
Christopherson, Sara, 107
Chu, Man-Wai, 025
Church, Jacqueline, 079
Circi, Ruhan, 152, 178
Cisterna, Dante
Cizek, Gregory, 121, 173
Clark, Amy, 154
Clausen, Mary Kate
Clauser, Brian, 093, 126, 178
Clauser, Jerome, 116
Clin Deffarges, Alexandre, 074
Close, Catherine, 038
Coates, Adam, 168
Cohen, Allan, 180, 048, 096
Cole, Veronica, 071
Colvin, Kimberly, 098, 122, 160
Confrey, Jere, 072
Cook, Linda, 113, 177
Copeland, Nicholas, 051
Copur-Gencturk, Yasemin, 180
Corliss, Mary-Alice
Corter, James, 075
Cortes, Juan, 144
Costello, De'Ani, 150
Courey, Karyssa, 059, 109
Crabtree, Ashleigh, 012
Crespo Cruz, Eduardo J, 061, 122
Crewe, Tara, 067
Croft, Michelle, 072
Cruce, Ty
Cruz, Alvaro
Cruz, Sandra, 144, 165
Cui, Mengyao, 081, 157
Cui, Weiwei, 073, 139
Cui, Wenju, 026, 062
Cui, Ying, 043, 114
Cui, Zhongmin, 157
Culpepper, Steven, 084, 170
Custer, Michael, 031
D'Brot, Juan, 098
D'sa, Nikhit
Dadey, Nathan, 110, 141, 186
Dadzie, Justice, 051, 123
Dai, Shenghai, 031, 085, 115, 182
Dailey, David, 157
Daisher, Ted, 033, 167
Dallas, Andrew, 031, 060, 146
Danielson, Robert, 031
Daoust, Carolyn, 180
Davidson, Anne, 098
Davis, Laurie, 148
Davis-Becker, Susan, 046, 138, 183, 190
de la Torre, Jimmy, 015, 054, 073, 122, 133, 148
De Lisle, Jerome, 188
de los Reyes, Andy, 068, 151
Deane, Paul, 036, 058
Debelak, Rudolf, 074
Debnam, Candace
DeCarlo, Lawrence, 104
Delaney, Nora, 061
Demars, Christine, 054, 103
Demir, Cihan, 090
Denbleyker, John
Deng, Jiayi, 073
Deng, Nina, 077
DePascale, Charles, 177
Deters, Lauren, 027
Di Digiacomo, Daniela, 172
Diao, Hongyu, 030
Dicerbo, Kristen, 027
Dickinson, Emily, 135
Diegel, Mary Helen, 009
Ding, Jiyuan, 054, 136
DiRenzo, Raegan, 181
DiStefano, Christine, 051
Dobria, Lidia, 122
Domingue, Benjamin, 028, 061, 111
Dong, Yixiao, 031, 122, 146
Doran, Harold, 024, 162
Douglas, Kerrie, 061
Drake, Philine, 132
Draney, Karen, 145
Dreyer, Stacy, 037, 181
Drost, Bryan, 035, 042, 067
Du, Ying, 178
Duan, Qizhou, 059
Dumas, Denis, 031, 146
Dunbar, Stephen, 052, 061, 103, 114
Dunn, Jennifer, 022, 138, 162
Duran, Lillian, 105, 137
Durmus, Aybuke, 180
Dwyer, Andrew, 190
Dym, Abigail, 039
E. Tran, Jasmine, 111
Eagan, Brendan, 123
Eckerly, Carol
Eckes, Thomas, 084
Edeh, Ejike, 084
Edi, Daniel, 189
Edwards, Kelly, 190
Effatpanah Hesari, Farshad, 131
Egan, Karla, 175
Eggers-Robertson, Lisa, 178
Embretson, Susan, 083, 112, 181
Emery, Matt, 064

PARTICIPANT INDEX

(Last Name, First Name, Session Number)

Emler, Trina, 171
Engelhard, George, 103
Englert, Kerry, 105, 175
Ercikan, Kadriye, 113, 140
Erickson, Lori, 176
Escribano, Rosario
Espinosa Aguirre, María Jesús
Estrada, Samantha, 146
Evans, Carla, 055, 118, 155
Everson, Howard, 098
Eze, Jesse, 051
Ezugwu, Ifesinachi, 051
Fan, Fen, 093
Fan, Meng, 062, 063, 135
Fan, Ziyu, 100, 148
Fancsali, Steven, 186
Fang, Luyang, 112
Fauss, Michael, 101
Fay, Derek, 157
Federiakin, Denis, 122, 132
Feinberg, Richard, 022, 041, 051, 080, 157
Feng, Jing, 121
Feng, Mingyu, 072
Feng, Xiaoying, 153
Feng, Ying
Feng, Zechu
Ferrara, Steve, 055, 057
Ferris, Julia, 040
Fickes, Natalie, 067
Figueroa Millares, Javiera
Filonczuk, Audrey, 104
Fina, Anthony, 052, 073, 103
Finch, W. Holmes, 031
Fincher, Melissa, 035
Fink, Aron, 099
Finney, Sara, 133
Firoozi, Tahereh, 114
Fisk, Charles, 159
Fitzsimmons, Ellen, 034
Flor, Michael, 059
Fong, Michael, 053, 149
Forsyth, Carol, 072
Forte, Ellen, 035, 057, 068, 113, 138, 151, 185
Foster, Chris, 130
Foster, David, 070, 130
Foster, Paul, 034
Fraser, Antonio, 040, 072
Fremer, John, 088
French, Brian, 031, 085
Freund, Dave
Freund, Rebecca, 103
Frey, Andreas, 094, 099
Fu, Jianbin, 086
Fu, Yanyan, 054
Fuentes-Jimenez, Mia, 051, 061, 083, 137
Fulmore, Julia, 146
Furgol Castellano, Katherine, 051, 102
Gabler, Jennifer, 132
Gallagher, Melissa, 123
Gan, Kaiyu, 053
Gandara, Fernanda
Gandhi, Kanishk, 111
Gane, Brian, 123
Gao, Yizhu, 114, 176
Ge, Yuan, 105
Gearin, Brian, 176
Geisinger, Kurt, 044
Gerker, Heather, 171, 180
Ghonsooly, Behzad, 122
Gianopulos, Garron, 186
Gierl, Mark, 114, 122, 147
Giers, Benjamin, 025
Giers, David, 025
Gilbert, Joshua, 018, 028
Gill, Matthew, 029
Gilmore, Jack, 103
Gitiria, Lucy, 053, 168
Gkolia, Aikaterini, 124
Glore, Christy, 181, 185
Gochyyev, Perman, 145
Godoy, María Inés, 144
Golden, Richard, 161
Gonzalez, Mario, 146
Gooch, Reginald
Goodman, Joshua, 031, 094, 138
Goodman, Noah, 111
Goodwin, Sarah, 079
Gopalakrishnan, Madhumitha, 147
Gordon, E. Wyatt, 046
Gorgulu, Rumeysa, 042, 051
Gorgun, Guher, 147, 152
Gorin, Joanna, 037
Gorney, Kylie, 004, 078, 094, 116, 134
Gorno Tempini, Marilu, 105, 137
Gorum Bing, Crystal, 126
Gotzmann, Andrea, 076
Goulart, Sarah, 179
Graf, Edith, 072
Gregg, Nikole, 104, 146
Greiert, Brittany, 118
Grekov, Paulina, 049
Gribben, Monica, 110
Griger, Cassondra, 053
Grochowalski, Joe, 121
Gui, Yi
Guo, Enruo, 026
Guo, Hongwen, 054, 099
Guo, Shuchen, 164
Guo, Wenjing, 083
Guo, Yage, 083
Guo, Yanhui, 178
Gupta, Vishal
Gushta, Matthew, 086, 137
Haab, Sergio, 156
Haber, Nick, 111
Habing, Brian, 061
Hadi, Syed Abdul, 083, 129, 157
Halai, Anjum, 188
Halpin, Peter, 092, 119
Ham, Dae Woong, 063
Ham, Eun Hye, 063
Hamdani, Maria, 117
Hamilton, Laura, 039, 068, 118, 151
Hammami, Sarah, 103
Han, Kyung (Chris), 052, 054, 083, 094, 137, 166
Han, Suhwa, 077, 086, 123, 135, 147
Han, Youngjin, 158
Han, Yuting, 053
Hao, Jiangang, 026, 062
Hao, Kevin, 108
Harik, Polina, 093, 178
Harra, Kjorte, 168
Harrold, Archer, 023
Harrold, Brian, 061
Hartig, Johannes, 132
Hass, Christopher, 025
Hatch, Amber, 114
Hauenstein, Clifford, 083
Haviland, Sara
He, Joy, 111
He, Qiwei, 016, 043, 051, 059, 068, 092, 112, 151, 157
He, Surina, 043
He, Yi, 100
Headley, Mollie, 176
Hebert, Andrea, 080
Heckler, Andrew, 032

PARTICIPANT INDEX

(Last Name, First Name, Session Number)

Hefty, Tyler, 029
Hellman, Scott, 040, 090
Hembry, Tracey, 100
Hemphill, Cadelle
Henderek, Jerusha, 093
Hendrickson, Amy, 121
Henríquez, Catalina, 144
Henson, Robert, 136
Hernandez, Jasmin, 047
Hernández, Diley, 073
Herr, Riley, 133
Herrada, Maximiliano, 144
Hicks, Lloyd
Himelfarb, Igor, 083, 124
Ho, Andrew, 088, 177
Ho, TsungHan, 083
Hodes, Paul, 132
Hodge, Kari, 146, 178, 192
Hodge-Clark, Kristen, 109
Hoeve, Karen, 178
Hoffman, Alexander, 019, 099, 181, 185
Hojnoski, Robin, 073
Holder, Eli, 102
Holliday, Keith, 167
Hong, Hyeri, 127, 160
Hong, IkHyeon, 114
Hong, Seong Eun, 083
Hooks-Singletary, Brianna, 069
Hopfenbeck, Therese, 188
Horton, Omaya, 189
Hoskins, Jessica, 047, 098
Hou, Tiantian, 168
Houang, Richard, 103
Houriet, Andrew, 126
Hove, Jennifer, 025
Hsiao, I-Chun Vera, 061
Hsieh, Hsiaolin, 128
Hu, Sara, 047
Hu, Yian, 112
Hua, Cheng, 174
Hua, Haowei, 074
Huang, Anne, 051
Huang, Chi-Yu, 037
Huang, Jing, 187
Huang, Mingya, 043
Huang, Qi, 028
Huang, Sijia, 034, 092, 149
Huang, Wei, 085
Huang, Yingshi, 143
Huang, Yuan, 169
Huang, Yue, 040, 062, 178
Huff, Kristen, 057, 068, 138, 151
Huggins-Manley, Corinne, 181
Hughes, Sarah, 090, 169
Hunt, Jessica, 123
Hwang, Heungsun, 112
Hwang, Kyu-Ryng, 110
Hwang, Yaurie, 178
Iaccarino, Stephanie, 178
Ignacio, Angelie, 146
Ihlenfeldt, Samuel, 078, 107
Im, Sukkeun, 159
Immanuel, Aria, 051
Immekus, Jason, 031
Ing, Marsha, 039
Ingrisone, James, 129
Ingrisone, Soo, 129
Isbell, Daniel, 134
Ivanova, Militsa, 133
Iyer, Balaji, 181
Jackson Haynes, Janine, 069
Jahan, Israt, 025
Jain, Vipul, 093
Jang, Eunice, 146
Jarr, Karoline, 012
Jeffries, Jay, 061, 158
Jensen, Bryant, 097
Jeon, Minjeong, 143, 161
Jewsbury, Paul, 082, 101, 171
Ji, Feng, 053
Ji, Jeongmin, 061, 103
Ji, Xuejun, 031, 178
Jia, Ao, 122
Jia, Helena, 049
Jiang, Yang, 026, 072
Jiang, Yanlin, 031, 062, 063, 146
Jiang, Yanming, 073
Jiao, Hong, 048, 058, 074, 090, 101, 164
Jimenez, Auburn, 042, 170
Jimenez, Daniela, 144
Jimenez-Silva, Margarita, 122
Jin, Kuan-Yu, 031, 054, 084
Jin, Rong
Jin, Ying
Johanningmeier, Scott, 037, 181
Johnson, Evelyn, 064, 146, 189
Johnson, Matthew, 091, 099, 101, 140, 143
Johnson, Matthew, 006
Johnson, Odis, 140
Jones, Andrew, 093
Joo, Joanne, 061
Joo, Sean, 168
Joshi, Malt, 031
Ju, Unhee, 166, 174
Julian, Marc, 183
Jung, Ae Kyong, 156
Jung, Juyoung, 106
Kaçak, Tugay, 122
Kahne, Joseph, 039
Kamata, Akihito, 032, 034, 053, 062, 149, 168
Kane, Michael, 178
Kang, Hyeon-Ah, 077, 136
Kang, Youngsoon, 178
Kannan, Priya, 022, 102, 117
Kanopka, Klint, 028, 126, 169
Kantor, Cayce, 122
Kaplan, David, 168
Kapoor, Shalini, 037, 086
Kaptur, Bradley, 178
Kaptur, Danielle (Dandan), 073, 178
Kara, Yusuf, 032, 053, 062, 149, 168
Karamese, Hacer, 083, 129, 157
Karelitz, Tzur, 178
Karimi Malekabadi, Farzan, 135
Karvonen, Meagan, 154
Kary, David, 159
Kehat, Gitit, 184
Kehinde, Olasunkanmi, 085
Kelleigh, Willow, 178
Keller, Brian, 007
Keller, Lisa, 031, 073, 083, 130
Keller, Rob, 080
Keng, Leslie, 091
Kennedy, Patrick, 061, 176
Kern, Justin, 061, 073, 084, 136
Kerzabi, Emily, 026, 062, 102
Ketterlin Geller, Leanne, 061
Keum, Eunhee, 033
Kevelson, Marisol, 051, 102
Khurana, Havisha, 061
Kılıç, Abdullah, 122
Killion, Sarah, 118
Kim, Doyoung
Kim, Edward, 149
Kim, Eunbee, 083
Kim, Hahyeong, 157
Kim, Hyunah, 025
Kim, Hyunjoo, 122
Kim, Jieun, 070, 134, 157

PARTICIPANT INDEX

(Last Name, First Name, Session Number)

Kim, JongPil, 002, 031, 146
Kim, Kyung Yong, 106
Kim, Meereem, 094
Kim, Minjung, 171
Kim, Nana, 143
Kim, Seongeun, 106
Kim, Stella, 031, 073, 106, 130, 146, 159
Kim, Sungyeun, 031
Kim, Yejin, 103
Kim, YJ, 031
Kim, Young Yee, 153
Kim, YoungKoung, 105, 142, 184
Kim, Youngwon, 135
Kim, Yun-Kyung, 033, 063
King, Emma, 153
Kinsey, Devon
Kingston, Neal, 138
Kintz, Tara, 009, 180
Kirakosian, Antranik, 165
Kislyonkova, Alexandra, 104
Klein, Trish, 154
Koehn, Hans, 122
Kohmer, Anika, 132
Kokolis, Gerry
Kong, Lingchen, 093
König, Christoph, 099
Konstantinidou, Evi, 124
Koo, Miryeong, 063
Koval, Jayma, 073
Kozikowski, Andrzej, 031
Kremmel, Benjamin, 134
Krenzke, Tom, 049
Krist, Andrew, 110, 128
Kroopnick, Marc
Krost, Kevin, 053
Krumm, Andrew, 063
Krupar, Allyson
Kuang, Huan, 048, 087
Kukea Shultz, Pohai, 068, 151, 175
Kulczycki, Christy, 029
Kuncel, Nathan, 068, 151
Kunina-Habenicht, Olga, 131
Kunz, Ann-Kathrin, 132
Kuo, Tzu-Chun, 062, 064, 077
Kush, Joseph, 087, 049, 158, 168
Kwako, Alexander, 062, 090, 164, 184
Kwon, Muwon, 158
Kwon, Sunbeom, 031, 101
Kyllonen, Patrick, 026, 062, 086
Kyung, Jungwon, 114, 130, 181
Ladson-Billings, Gloria, 088
Laesecke, Anne
LaFlair, Geoff, 108, 134
Lahoud, Tamlyn, 096, 180
Lai, Hok Chio, 122, 191
Lai, Viet, 135
Laitusis, Cara, 025, 151
Lakin, Joni, 085, 128
Lan, Andrew, 060, 092
Lane, John, 009, 092
Langi, Meredith, 106, 176
Lavigne, Cheryl
Le, Anh Thu, 033, 059, 158
LeBeau, Brandon, 038, 106
Lederman, Josh, 073
Lee, Chansoon (Danielle), 061, 116, 159
Lee, Haeju, 149
Lee, Hyeryung, 129, 166
Lee, Hyunjung, 061, 075
Lee, Jeremy
Lee, Jiyoon, 157
Lee, Mina, 081
Lee, Minho, 161
Lee, Selene, 157
Lee, Sooyong, 083, 129, 157
Lee, Sung-Hyuck, 083, 094
Lee, Sunhyoung, 051
Lee, Won-Chan, 052, 106, 114, 130, 159, 164
Lee, Yongsang, 094, 114
Leeming, Meghan, 136
Lefstein, Adam, 097
Lehman, Blair, 124, 133
Lekwa, Adam, 127
Lembke, Erica, 061
Leng, Dihao, 054
Lentini, Maria, 157
Lerch, Ryan, 184
LeRoy, Stephanie, 103
Leventhal, Brian, 061, 087
Levi, Effi, 178
Lewis, Daniel, 057, 192
Lewis, Jennifer, 122
Lewis, John, 118
Lewis, Regina, 035, 046, 098
Li, Dongmei, 037, 127, 139
Li, Feiming, 048
Li, Grace, 154
Li, Ji, 075
Li, Jian, 168
Li, Jiawei, 063
Li, Jie, 187
Li, Jujia, 085
Li, Lanrong, 137
Li, Min, 114
Li, Ming, 101
Li, Shuhong, 171
Li, Shuyuan, 047
Li, Warren
Li, Wenshuo, 114
Li, Xiao, 059
Li, Xin, 029
Li, Yifei, 105
Li, Zhifei, 114
Liang, Min, 114
Liang, Xinya, 075
Liao, Manqian, 058, 108, 134
Liao, Xiangyi, 103
Liaw, Yuan-Ling
Lim, Hwanggyu, 094
Lim, Sangdon, 095, 121, 192
Lim, Youn Seon085, 170
Lin, Pei-Ying, 025
Lin, Ye, 146, 178, 192
Lindner, Marlit, 133
Ling, Guangming, 109, 126, 155
Lipner, Rebecca, 177
Litschwartz, Sophie, 120
Liu, Chunyan, 157
Liu, Hongyun, 053
Liu, Huan, 031
Liu, Jin, 073
Liu, Jinghua, 002, 106, 114, 139
Liu, Joyce Xinle, 048, 148
Liu, Junhui, 030
Liu, Lei, 062, 109
Liu, Lucia, 122, 157
Liu, Ou Lydia, 109, 177
Liu, Qiao, 130
Liu, Qidi, 048
Liu, Ran, 178
Liu, Ren, 104
Liu, Ruitao, 076
Liu, Xiaoxiao, 043, 114
Liu, Xinyu, 114
Liu, Yang, 158
Liu, Yaohui, 059
Liu, Ying, 170
Liu, Yunting, 111, 145
Lnu, Chandramani, 101
Lo, Kai-Ling, 108

PARTICIPANT INDEX

(Last Name, First Name, Session Number)

Lochbaum, Karen, 055
Lockspeiser, Tai, 125
Lockwood, J.R., 070, 108
Long, Haiying, 181
Long, Yiting, 049
López, Dany, 144
Lottridge, Susan, 036, 055, 074, 090, 092, 101, 123, 138, 162, 164, 184
Lou, Yaosheng, 122
Lovelace, Temple, 031
Lu, Ru, 171
Lu, Xinyi, 111
Lu, Yikai,095, 119, 147
Lu, Ying, 073, 083
Lu, Yining, 169
Luccioli, Viviana, 112
Luciw-Dubas, Ulana, 126
Ludovica, De Carolis, 143
Luecht, Richard, 017
Lugu, Benjamin Kweku, 157
Lui, Angela, 040, 096
Luo, Jinwen, 034, 143
Luo, Lin, 131
Lyons, Susan, 055, 107
Lyu, Weicong, 010, 119, 191
Ma, Jing, 052, 073
Ma, Wanjing (Anya), 028, 051, 061, 083, 111, 137
Ma, Wenchao, 015, 157
Ma, Ye, 014, 058, 089
Macias, Andrea, 047
MacRae, Heather, 081
Madison, Matthew, 083, 131
Maeda, Hotaka, 147
Maeda, Yukiko, 051, 127
Majerus, Mary, 042
Makinde, Henry, 060, 147, 168
Malatesta, Jaime, 052, 139, 187
Man, Kaiwen, 004, 059, 085, 171
Mann, Danielle, 070
Mao, Tianzheng, 061
Mao, Xia
Marder, Andrew, 130
Mardones-Segovia, Constanza, 077, 096, 180
Margolis, Melissa, 126, 177
María Leonor, Varas, 144
Marini, Jessica, 157, 179
Marion, Scott, 087, 097, 110, 141
Marquine, Maria, 068, 151
Martin de los Santos Kleinz, Lisa, 132
Martineau, Joseph, 045
Martinez, Alfonso, 156
Mason, Jessica, 090, 155
Mastrup, Kayce, 150
Mattern, Krista, 177
Maul, Andrew, 113
Maur, Andreas, 132
Maxfield, Maura, 122
Mazzullo, Elisabetta, 152
McBride, Yuanyuan, 095
McCaffrey, Daniel, 005, 040, 082, 101
McClure, Kyla, 141
McFadden, Mara, 051
McGee, Steven, 178
McGrane, Joshua, 113, 188
McMillan, James, 050, 067
McNamara, Danielle, 140
McVey, Jill, 091
McWilliams, Kara, 162
Mehou, Juste, 061
Mehrabi, Amirreza, 051, 168
Mehrazmay, Roghayeh, 122
Meijer, R.R., 168
Mena, Fernando, 165
Mendez, Jose, 047
Mendoza, Guadalupe, 047
Meng, Huijuan, 014, 056
Meng, Lionel, 103
Merkle, Ed, 034, 076
Merkt, Martin, 128
Mese, Can, 125
Meyer, Robert, 047, 172
Michaelides, Michalis, 133
Michaels, Hillary, 138
Michaluk, Lynette, 031
Michel, Rochelle, 068, 151
Middlestead, Andrew, 099
Middleton, Kyndra, 069, 150, 183
Miles, Julie, 121
Minchen, Nathan, 029
Ming, Norma, 109
Miratrix, Luke, 013, 063, 084
Mireles, Nicolas, 021, 114, 168
Mislevy, Robert, 003, 063
Missall, Kristen, 073
Mohammadi, Hamid
Mojoyinola, Mubarak, 029
Molerov, Dimitri, 132
Moore, Steven, 060
Morales, Amy, 091
Morales, María Cristina, 144
Morell, Linda, 066
Morphew, Jason, 051, 168, 187
Morrison, Kristin, 052
Moses, Tim, 076, 105, 106, 113, 142, 163
Moteane, Malitšitso, 069
Mottesi, Gabriela, 038
Moulder, Brad, 030
Mulcaire, Phoebe, 108
Muller, Eric, 153
Munevver, Ilgun Dibek, 180
Muntean, William, 093, 187
Murphy, Sarah, 083
Murray, Amelia, 180
Murray, Angela, 171, 180
Murray, Tonya, 178
Muszyński, Marek, 178
Myoung, Eunjung, 128
Nagasawa, Mirai, 041
Nagel, Marie, 132
Nagy, Gabriel, 133
Naim, Alper, 101
Nalbandyan, Roza, 051, 179, 189
Nam, Sungjin, 178
Nash, Brooke, 042, 170
Nason, Kelsey, 093, 169
Nathan, Rebecca
Nesbitt, Jaylin, 038, 053, 069
Nese, Joseph F. T., 032
Neumann, M., 168
Newton, Steve, 127
Ngo, Trung, 135
Nguyen, Hanh, 031
Nguyen, Hiep, 135
Nguyen, Tram-Anh Tran, 61
Nickodem, Kyle, 75
Niessen, A.S.M, 168
Nissen, Jayson, 187
Nitkin, David, 109
Nolan, Kate, 178
North, Kai, 092, 108
Novak, Jennifer, 178
Nydick, Steven, 092, 108
O'Donnell, Francis, 022
O'Grady, Christina, 168
O'Rourke, Kevin, 080
Oakley, Alexis, 103, 150
Ober, Teresa, 059, 74, 109, 190
Oberle, Catherine, 176

PARTICIPANT INDEX

(Last Name, First Name, Session Number)

Oberle, Kristina, 091
Obot, Patrick, 123
Odemwingie, Michelle, 107
Oh, Hyeonjoo, 002, 031, 061, 076, 083, 157
Oh, Hyunjee, 157
Oliva, Michelle, 172
Olivera Aguilar, Margarita, 039
Oliveri, Maria, 023, 051, 055, 061, 098
Olney, Andrew, 060
Olson, Jacinta, 123
Olson-Rickel, Cara, 185
Ong, Thai, 126
Oostrom, Janneke, 031
Ormerod, Christopher, 027, 036, 060, 079, 090, 0142, 152, 164, 184
Orrill, Chandra, 096
Osaro, Sarah, 114
Oyeniran, Daniel, 128
Ozkan, Fatih, 103
Palermo, Corey, 036, 062, 142
Pan, Qianqian, 191
Pan, Tianshu, 083, 146
Pan, Yanjun, 125
Panadero, Ernesto, 112
Panzer, Jakob, 114
Papanastasiou, Elena, 124
Pardos, Zachary, 111, 174
Park, Chan, 158
Park, Hyemin, 146
Park, Junhee, 061, 114
Park, Mi Jin, 141
Park, Nohwon, 180
Park, Seohee, 041
Park, Sojung, 122
Park, Yena, 079
Park, Yookyeong, 122
Parker, Emily, 064, 146, 189
Pastor, Dena, 133
Patelis, Thanos, 155, 181
Patterson, Chris, 146
Patterson, Jaylene, 031
Patterson, Luke, 178
Patz, Richard, 020, 040
Pei, Bo, 059
Pei, Francesca, 137
Pek, Jolynn, 032
Perie, Marianne, 078
Perkoff, Elana Margaret, 178
Perna, Megan, 176
Pezeshki, Maryam, 042, 095, 186
Pham, Duy, 067, 135
Phelps, Geoffrey, 190
Phenow, Aurore, 061
Phillips, Gary, 049, 121
Pilner Blair, Kristen, 150
Pinto, Wallace, 150
Pitoniak, Mary, 113, 177
Plisiecki, Hubert, 178
Poe, Mya, 023, 055, 098, 173
Pokropek, Artur, 178
Polikoff, Morgan, 177
Poole, Frederick, 178
Por, Han-Hui
Porcelli, Victor, 008
Potgieter, Cornelis, 034
Powell, Sarah, 061
Powers, Sonya, 033, 042, 095, 100, 135, 176
Prendergast, Caroline, 093
Prewett, Sara, 064, 146
Proctor, Thomas, 058, 070
Purpura, David, 073
Pusey, Kerry, 175
Qiao, Xin, 034, 048, 053
Qin, Qi, 053, 099, 174
Qiu, Chen, 037
Quanbeck, Mari, 066
Quesen, Sarah, 002, 038, 055, 176
Quirk, Victoria, 084
Rai, Snigdha, 077
Rajeb, Mehdi, 128, 157
Ramazan, Onur, 031, 099, 128
Rambis, Natalie, 098
Ranieri, Linda, 126
Ravand, Hamdollah, 131
Rayfield, Norah, 025
Razavi, Pooya, 042, 100, 135
Read, Daniel, 118
Read, Wyatt, 180
Reckase, Mark, 045, 094, 153, 166, 174
Ren, He, 170, 191
Rewley, Kelly, 116
Rezayi, Saed, 093
Richie-Halford, Adam, 061, 083
Rijmen, Frank, 081, 123, 145, 147
Rikoon, Samuel, 039
Riley-Lepo, Erin, 067
Rios, Joseph, 091, 150
Rios, Kristina, 160
Rios, Oscar, 122
Ritter, Steven, 186
Ritthaler, Mike, 090
Roberts, Kari, 114
Robertson, Maya, 165
Robin, Frederic
Robitaille, Jennifer, 064, 146, 189
Rodriguez, Jennifer, 150
Rodriguez, Michael, 020, 047, 068, 075, 103, 138, 151, 189
Rodríguez, Nelson Andrés, 032
Roeber, Edward, 009
Roeters-Solano, Heather, 055
Rolla, Andrea, 038
Rome, Logan, 033, 167
Rosales de Veliz, Leslie Vanessa
Rosca, Oxana, 096
Roussos, Louis, 053, 061, 099, 129, 174
Rozman, Mojca, 064
Ruiz-Primo, Maria, 050, 128
Ruiz Diaz, Shona, 072
Runge, Andrew, 079, 108
Runyon, Christopher, 041
Rupp, Andre, 085
Russell, Javarro
Russell, Michael, 068, 098, 151, 154
Rutkowski, Leslie, 082, 115
Rutzler, Erica
Ryan, Ève, 168
Saavedra, Ana, 061, 083, 137
Saavedra, Cristian, 144
Sabatini, Peter, 157
Sachdeva, Ruchi, 188
Sahakyan, Narék, 105
Şahin Kürşad, Merve, 180
Saiar, Amin, 138
Sajidin, Zahra, 157
Sakovich, Jeremy, 159
Saldivia, Luis, 099, 165
Salzer, Rebecca, 008
Samant, Yashad, 090
Sanchez, Edgar, 125, 179
Sanchez, Sergio, 150
Sandoval Hernandez, Andres, 103
Santelices, Maria, 179
Sarac, Merve, 116, 121
Sarig, Joni, 178
Satkus, Paulius, 137, 159
Sato, Edynn, 154, 175
Sauder, Derek, 126

PARTICIPANT INDEX

(Last Name, First Name, Session Number)

Savage, Corey, 039
Savoie, Renee, 102
Sayin, Ayfer, 122, 147
Schellman, Madeline, 063, 170
Schenke, Katerina, 047
Schlomer, Gabriel, 122
Schmidt, Katie, 099, 174
Schneider, Christina, 029, 057, 121, 192
Schneider, Wei, 086
Schröter, Hannes, 128
Schumacker, Randall, 083
Schwartz, Bob, 125
Sciolla, Bruna, 144
Seissus, Emilia, 144
Serrentino, Dean, 167
Sgammato, Adrienne
Shaffer, David, 123
Shamsaldeen, Fajer, 073, 157
Shanahan, Marie-Claire, 025
Sharpnack, James, 108
Shaw, Emily, 157, 179
Shear, Benjamin, 065, 180
Shen, Yawei, 030
Shenavai, Kevin, 132
Shepard, Lorrie, 097
Sheppard Miller, Gayle, 175
Sherer, David, 109
Shermis, Mark, 036, 142, 164
Shi, Qingzhou, 157
Shillo, Roi, 178
Shin, David, 073
Shin, Emily, 146
Shin, Jihye, 146
Shin, MinJeong, 157
Shin, Seungwon, 130
Shor, Yael, 178
Shukla, Kathan
Shyyan, Vitaliy, 175
Siebert, Julian, 032, 061, 083, 105, 137
Sikali, Emmanuel, 178
Simmering, Vanessa, 154
Simpson, mary Ann, 186
Sinharay, Sandip, 078, 080, 091, 119
Sireci, Stephen, 046, 061, 068, 088, 108, 151, 173
Sirganci, Gozde, 125, 149
Siu, Amanda, 181
Skerrett, Allison, 163
Skorupski, William, 171
Slater, Sharon, 164
Sloan, Caroline, 185
Small, Electra, 008
Smith, Jessalyn, 078, 174, 191
Smith, Nate, 080
Snow, Erica, 064
Soland, James, 071, 103
Solano-Flores, Guillermo, 050, 113, 128
Somay, Su, 126
Someshwar, Shonai, 029
Somsong, Sarunya, 053
Sondermann, Christina, 128
Song, Dan, 061, 074, 114, 164, 178
Song, Victoria, 166
Song, Woori, 103
Song, Yi
Soni, Parth, 065
Sonnleitner, Philipp, 131
Southerland, Joshua, 040, 090
Soyoye, Olushola, 051
Sparks, Jesse, 124, 133
Specht-Boardman, Ryan, 155
Stacho, Ladislav, 072
Stahlhut, Laura, 074
Stark, Stephen, 068, 151
Steedle, Jeffrey, 029, 057, 091
Stegers-Jager, Karen, 031
Stethen, Kelley, 114, 190
Stevens, Mark, 134
Stirn, Yuanyuan, 052
Stout, William, 136
Strietholt, Rolf, 103
Student, Sanford, 033, 053, 103, 120, 167, 180
Su, Yu, 176
Su, Yu-Lan, 125
Subhiyah, Raja, 157
Suh, Hongwook, 146
Suh, Yon Soo, 104, 161
Suk, Youmi, 158
Sun, Sundance Zhihong, 064
Sun, Yulan, 144
Sundar, Kripa, 047
Sung, Youjin, 158
Sunnassee, Devdass, 089
Suolang, Deji, 115
Supovitz, Jonathan, 073
Surman, Stacy, 051
Svetina Valdivia, Dubravka, 103, 149, 168
Swaby, Jeneve, 043, 115
Sweeney, Sandra, 129
Sweet, Tracy, 048
Swinburne Romine, Russell, 154
Swinth, Emily, 008
Szendey, Olivia, 038
Tajik, Elham, 063
Talreja, Vinita, 014
Tan, Rachael, 125
Tang, Cheng, 048, 096, 103, 126
Tang, Nai-En, 083, 146
Tang, XiaoDan, 051
Tang, Xiuxiu, 136
Tang, Yuefeng, 073
Tao, Shuqin, 042, 095, 186
Taub, Deborah, 066
Tavares, Stephen, 071
Taylor, Catherine, 055
Taylor, Christopher, 014
Taylor, Melinda, 029
Templin, Jonathan, 071, 123, 146, 156
Tenison, Caitlin, 190
Terao, Takahiro, 086
Thacker, Arthur, 110
Thier, Michael, 039
Thomas, Jay, 037, 181
Thomas Murphy, Ciara, 150
Thompson, Jake, 042, 161, 170
Thompson, Nathan, 166
Thum, Yeow Meng, 033, 065
Thurlow, Martha, 066
Tighe, Elizabeth, 016
Tingir, Seyfullah, 127
Toker, Turker, 131
Tolan, Shari, 185
Tolentino, Lissette, 179
Tong, Ye, 068, 091, 138, 151, 167, 183
Topczewski, Anna, 159, 183
Torres Irribarra, David, 144, 179
Toton, Sarah, 004
Townley-Flores, Carrie, 051, 061, 083, 137
Toyama, Yukie, 145
Tran, Jasmine, 051
Tran, Peter, 061
Trierweiler, Lukas, 132
Trout, Nick, 116
True, Taylor, 064
Tsai, Chia-Lin, 146
Tse, Winnie Wing-Yee, 191

PARTICIPANT INDEX

(Last Name, First Name, Session Number)

Tu, Dongbo, 054, 104, 122, 136
Tucker, Eric, 039
Turner, Ronna, 073, 089, 114
Twing, Jon, 046, 128
Ulitzsch, Esther, 133
Umeh, Emmanuel, 123
Ungashe, Albu, 051
Vafis, Greg, 039
Valdivia Medinaceli, Montserrat, 091, 148
van Buel, Sanjay, 031
van Campenhout, Rachel, 060
Van Dusen, Ben, 187
Van Orman, Dustin, 050, 067, 182
Vang, Gao, 176
Velez, Sophia, 172
Ventura, Claudia, 121
Vernon, Annette, 041, 061, 103
Villafuerte, Catherina, 021, 098
Villanueva, Kristin, 130
Villegas, Leslie, 102
Vispoel, Walter, 129, 160, 191
Vivas Corrales, Pablo, 103
Vo, Thao, 117
Vo, Yen, 146
von Davier, Alina, 017, 058, 070, 092, 108, 162, 166
von Davier, Matthias, 061, 082, 124, 178
Vorenkamp, Ellen, 009
Vu, Lien, 103
Vue, Kory, 176
Wackerle-Hollman, Alisha, 176
Walker, Marcus, 060
Walker, Michael, 098, 138, 173, 183
Walkington, Candace
Walkowiak, Temple
Wallace, Leslie
Wan, Lei, 083, 121
Wan, Siyu, 030
Wang, Aijun, 178
Wang, Bowen, 041, 048
Wang, Caroline, 047
Wang, Chun, 010, 028, 119, 170, 186
Wang, Chunxin, 076
Wang, Daxun, 054, 122
Wang, Dongwei, 030
Wang, Fang, 084
Wang, Haiyan, 157
Wang, Hongling, 037
Wang, Huijuan, 051, 174
Wang, Kun, 096
Wang, Kuo, 062, 074
Wang, Min, 083, 139
Wang, Qin, 136
Wang, Shaojie, 034
Wang, Shichao, 076
Wang, Shiyu, 048, 096, 112, 161, 180
Wang, Tao, 084
Wang, Ting, 093, 178
Wang, Xiyu, 127
Wang, Xu, 111
Wang, Yu, 147
Wang, Yuan
Wang, Ze, 014
Wang, Zhen, 170
Wang, Zhuoran, 187
Wang, Zifan, 142
Wang, Zuowei, 058, 163
Warner, Jeffrey, 109
Warner, Zachary, 035, 138
Watts, Field
Webb, Blue, 178
Webb, Noreen, 026
Weeks, Jonathan, 045, 058, 075, 094
Wei, Elizabeth, 095
Wei, Hsin-Ro, 083
Wei, Junhuan, 054, 104, 136
Wei, Xiaobo, 084
Wei, Xin, 016, 025
Weinstein, Marc, 075
Weiss, David, 058, 166
Welch, Catherine, 061, 103, 114, 173
Wells, Craig, 103, 114, 130
Welsh, Megan, 065, 150
Wen, Hongbo, 114
Wentzlof, Kelly, 051
Westrick, Paul, 157, 179
Wheeler, Jordan, 048
Whitcomb, David, 185
White, Lauren, 038, 055, 098
Whitmer, John, 027
Wilberforce, Winifred, 032
Wild, Autumn, 087, 168
Wiley, Andrew, 138
Will, Kelsey, 176
Williams, Angie, 168
Williams, Kevin, 109
Williamson, Benjamin, 137
Wilson, Joshua, 036, 062, 142
Wilson, Mark, 040, 068, 096, 145, 151, 177
Wilson, Quinell, 074
Wind, Stefanie, 041, 083, 128
Wine, Marjorie, 019, 099, 181, 185
Winter, Sonja, 007, 034
Wise, Steven, 133
Witmer, Sara, 114
Wolfe, Edward, 078, 090, 101
Wollack, James, 056
Wong, George, 157
Woo, Ada, 138
Wood, Scott, 037, 062, 116, 169
Woolf, Sherri, 184
Worrell, Frank, 068, 151
Worthington, Michelle, 099
Woulfin, Sarah, 097
Wright, Daniel, 114
Wright, Keith
Wry, Erin, 178
Wu, Meng
Wu, Shushan, 112
Wu, Sirui, 051
Wu, Tong, 061, 130, 146, 157
Wu, Xiaopeng, 131
Wu, Yi-Chen, 066, 082
Wu, Yi-Fang, 062, 085, 157, 192
Wu, Zebing, 086
Wyatt, Jeff, 121
Wylen, Mel, 038
Wylie, E. Caroline, 039
Wyman, Austin, 168
Wyse, Adam, 029, 052
Xi, Nuo
Xia, Yan, 083
Xiao, Daihui, 129
Xiao, Min, 096
Xiao, Xingyao, 040, 145
Xiong, Jiawei, 030, 042, 048, 054, 103
Xiong, Yao, 064
Xu, Gongjun, 010, 119
Xu, Guoliang, 075
Xu, Jiajun, 086, 157
Xu, Jing-Ru, 157
Xu, Rujun, 103
Xu, Xiaochen, 073, 122
Xu, Yangmeng, 078, 096
Xu, Yingshi, 114
Xue, Mingfeng, 062, 096, 147, 191
Yan, Duanli, 003, 017, 056, 058, 072, 166

PARTICIPANT INDEX

(Last Name, First Name, Session Number)

Yan, Pengfei, 074
Yancey, Kevin, 108
Yaneva, Victoria, 060, 093, 126
Yang, Ji Seung, 158
Yang, Yaxuan, 096
Yang, Yi, 030, 126
Yao, Lihua, 051
Yarmohammadtoosky, Sahar, 093
Yavuz, Sinan, 049, 115
Ye, Sangbeak, 015, 073, 148
Yeatman, Jason, 051, 061, 083, 111, 137
Yeboah, Clement, 051
Yee, Darrick, 032
Yen, Shu Jing, 083
Yin, Yue, 157
Yoo, Hanwook, 002, 076, 122
Young, Mackenzie, 074, 184
Young, Rebecca
Yu, Haoyang, 103, 168
Yu, Kefan, 114
Yuan, Ye, 094
Yung, Sammy, 025
Zahner, Doris, 039
Zapata-Rivera, Diego, 003, 022, 072, 169
Zapparrata, Nicole, 192
Zegers, Mónica, 105, 137
Zeiser, Kristina, 109
Zelikman, Eric, 111
Zeng, Biao, 143
Zeng, Ji, 099
Zenisky, April, 022, 122
Zhai, Xiaoming, 164
Zhang, Bo, 083
Zhang, Chuqing, 181
Zhang, Fan, 142
Zhang, Jihong, 075, 156
Zhang, Jinming, 063, 073
Zhang, Lijin, 028
Zhang, Mingqin, 014
Zhang, Nan, 101
Zhang, Oliver, 058, 166
Zhang, Susu, 031, 048
Zhang, Tiejun, 084
Zhang, Ting, 093, 178
Zhang, Wei, 073
Zhang, Xiaohang, 051
Zhang, Xiuyuan, 073, 083
Zhang, Yichi, 121
Zhang, Yifan, 124, 168
Zhang, Yu, 178
Zhang, Yuxiao, 127, 187, 191
Zhang, Zhiyong, 168
Zhao, Xinchu, 064
Zheng, Guoguo, 127
Zheng, Xiaying, 115, 153
Zheng, Yi, 092
Zhong, Xiaoting, 115
Zhou, Tianyi, 101
Zhou, Xinchang, 083
Zhou, Xuechun, 122, 157
Zhou, Yan, 135
Zhou, Yiyun, 093, 178
Zhu, Sizheng, 064
Ziedzor, Reginald, 093, 178
Zlatkin-Troitschanskaia, Olga, 122, 132
Żółtak, Tomasz, 178
Zopluoglu, Cengiz, 137
Zou, Tongtong, 116
Zu, Jiyun, 152
Zubenko, Vladimir, 101
Zwick, Rebecca, 044, 113

PARTICIPANT EMAIL

(Last Name, First Name, Affiliation, Email)

Ketan, University of Massachusetts, Amherst, ketan@umass.edu
Şahin Kürşad, Merve, TED University, merve.kursad@tedu.edu.tr
Żółtak, Tomasz, IFiS PAN, tomasz.zoltak@ifispan.edu.pl
Abeysinghe, Bhashithe, American Institutes for Research, babeysinghe@air.org
Abroampah, Doris, University of Alberta, abroampa@ualberta.ca
Abulela, Mohammed, University of Minnesota, mhady001@umn.edu
Acharya Julien, Nisha, Center for Measurement Justice, nacharya@measurementjustice.org
Ackerman, Terry, University of Iowa, terryackerman2@gmail.com
Adegoke, Hope, University of North Carolina, hopeadegoke13@gmail.com
Ahadi, Stephan, Cambium Assessment Inc., stephan.ahadi@cambiumassessment.com
Aiello, Liam, University of California, Berkeley, laiello@berkeley.edu
Akande, Christiana, National Commission on Certification of Physician Assistants, christianaa@nccpa.net
Akbay, Lokman, Department of Educational Sciences, Istanbul University-Cerrahpasa, lokmanakbay@iuc.edu.tr
Alahmadi, Sarah, ACS Ventures, LLC, salahmadi.research@gmail.com
Alarcón, Eduardo, DEMRE - Universidad de Chile, eduardoalarcon@uchile.cl
Albano, Tony, University of California - Davis, adalbano@ucdavis.edu
Alila, Nancy, Department of Educational Psychology, University of Georgia, Athens, GA, USA, nancy.alila@uga.edu
Allen, Jeff, ACT, Inc., jeff.allen@act.org
Almond, Russell, Florida State University, ralmond@fsu.edu
Amoateng, Ernest, Western Michigan University, ernestyaw.amoateng@wmich.edu
An, Lily, lily_an@g.harvard.edu
Anderson, Elizabeth, Founder, Embraced Wisdom Resource Group, eanderson@embracedwisdom.com
Andrade, Heidi, University of Albany, handrade@albany.edu
Andrade-Lotero, Luis Alejandro, Pearson, alejandro.andrade.lotero@pearson.com
Andrews-Todd, Jessica, ETS, jandrewstodd@ets.org
Anghel, Ella, Boston College, anghel@bc.edu
Anguiano-Carrasco, Cristina, ACT, cristina.Anguiano-Carrasco@act.org
Araneda, Sergio, Caveon, sergio.araneda@caveon.com
Arboleda, Juan, j.c.arboleda.cardona@rug.nl
Arias, Angel, Carleton University, AngelArias@cunet.carleton.ca
Armoush, Nour, University of Alberta, armoush@ualberta.ca
Arslan, Burcu, ETS Research Institute, BARSLAN@ets.org
Arthur, Ann, ACT Education Corp., Ann.arthur@act.org
Arthur, David, University of Washington, dbarthur@uw.edu
Asamoah, Nana Amma, University of Arkansas, nbasamoa@uark.edu
Asare, Eric, Old Dominion University, asareeric12@gmail.com
Athanases, Steven, University of California, Davis, szathanases@ucdavis.edu
Attali, Yigal, yigal@duolingo.com
Au, Wayne, University of Washington Bothell, wayneau@uw.edu
Augustsson, Keith, Pearson, keith.augustsson@pearson.com
Avila, Natalia, Pontificia Universidad Catolica de Chile, Facultad de Educación, naavila@uc.cl
Avraam, Demetris, University of Copenhagen, avraam.demetris@gmail.com
Ayers-Wright, Elizabeth, Cambium Assessment, eaa18sb@gmail.com
Badrinarayan, Aneesha, Education First, aneesha.badrinarayan@gmail.com
Bahry, Louise, louise.bahry@gmail.com
Bailey, Alison, UCLA, abailey@gseis.ucla.edu
Bailey, Jennifer, Educational Testing Service, jbailey004@ets.org
Bailey, Paul, American Institutes For Research, pbailey@air.org
Baker, Doris Luft, The University of Texas at Austin, doris.baker@austin.utexas.edu
Baker, Eva, University of California, Los Angeles, baker@cse.ucla.edu
Baldwin, Peter, National Board of Medical Examiners, pbaldwin@nbme.org

Balisciano, Nicholas, Harvard University, nbalisciano@alumni.upenn.edu
Bamat, David, American Institutes for Research, dbamat@air.org
Banjanovic, Erin, erin.banjanovic@gmail.com
Bao, Yu, James Madison University, bao2yx@jmu.edu
Baral, Kushmakar, University of Denver, Kushmakar.Baral@du.edu
Barber, Justin, Pearson, justin.barber@pearson.com
Barnes, Sophie, Harvard Graduate School of Education, spb702@mail.harvard.edu
Barnum, Stuart, The National Board of Osteopathic Medical Examiners, SBarnum@nbome.org
Baron, Zach, James Madison University, baronzd@dukes.jmu.edu
Barry, Carol, American Board of Surgery, cbarry@absurgery.org
Bates, Shaun, Missouri Department of Elementary and Secondary Education, shaun.bates@dese.mo.gov
Bay, Luz, College Board, lbay@collegeboard.org
Beach, Sarah, University of Virginia, seb3mr@virginia.edu
Beard, Jonathan, College Board, jonathan.j.beard@gmail.com
Bechard, Sue, ATLAS, University of Kansas, sue.bechard@ku.edu
Becker, Kirk, Pearson VUE, kirk.becker@pearson.com
Becker, Lee, Pearson, lee.becker@pearson.com
Bediwy, Ahmed, The University of Iowa, ahmed-bediwy@uiowa.edu
Behan, Bradley, Pearson, bradley.behan@pearson.com
Beiting-Parrish, Magdalen, Federation of American Scientists, magdalen.beiting@gmail.com
Bell, Courtney, courtney.bell@wisc.edu
Bellows, Martha, Pearson, martha.bellows@pearson.com
Belov, Dmitry, Law School Admission Council, dbelov@lsac.org
Belur, Vinetha, ETS, vbelur@ets.org
Belzak, William, Duolingo, wbelzak@gmail.com
Bennett, Randy, ETS, randyebennett@yahoo.com
Bergner, Yoav, New York University, yoav.bergner@nyu.edu
Bernacki, Matthew, University of North Carolina at Chapel Hill, mlb@unc.edu
Bernard, Steve, Luxembourg Centre for Educational Testing, University of Luxembourg, steve.bernard@uni.lu
Beswick, Bruce, The University of Melbourne, bruceab@unimelb.edu.au
Betts, Joe, NCSBN, jbetts@ncsbn.org
Beuoy, Aaron, University of Kansas, abeuoy@ku.edu
Bezirhan, Ummugul, Boston College, bezirhan@bc.edu
Bhadra, Dhiman, Indian Institute of Management Ahmedabad, dhiman@iima.ac.in
Bhandari, Shreya, University of California, Berkeley, shreya.bhandari@berkeley.edu
Biancarosa, Gina, University of Oregon, Ginab@uoregon.edu
Bilgic-Erdem, Serife, William & Mary, serifebilgic7@gmail.com
Binzak, John, Pearson Clinical Assessments, john.binzak@pearson.com
Bishop, Kyoungwon,kei.bishop2@gmail.com
Bishop, Norman, LSAC, n_scott_bishop@yahoo.com
Bohrnstedt, George, American Institutes for Research, GBohrnstedt@air.org
Bolender, Brad, Finetune by Prometric, brad.bolender@gmail.com
Bolt, Daniel, University of Wisconsin - Madison, dmbolt@wisc.edu
Bonge, Nicole, University of Arkansas, ngbonge@uark.edu
Bonifay, Wes, University of Missouri, bonifayw@missouri.edu
Bonk, Bill, Pearson, william.bonk@pearson.com
Bonner, Sarah, Hunter College, CUNY, sbonner@hunter.cuny.edu
Born, Marise, Erasmus University Rotterdam, m.ph.born@essb.eur.nl
Botter, Preston, pdbotter@gmail.com
Boughton, Keith, Data Recognition Corporation, kboughton@datarecognitioncorp.com
Bovaird, James, University of Nebraska-Lincoln, jbovaird2@unl.edu
Boyd, Aimee, aimeeboyd@cainc.com
Boyd, Brittany, American Institutes for Research, bboyd@air.org
Boykin, Allison, Cambium Assessment, allison.boykin@cambiumassessment.com
Boykin, Malik, Brown University, c_boykin@brown.edu

Bradley, Kelly, University of Kentucky, kelly.bradley@uky.edu
Bradshaw, Laine, Pearson, lainebradshaw@gmail.com
Braun, Henry, Boston College, braunh@bc.edu
Brennan, Robert, The University of Iowa, robert-brennan@uiowa.edu
Brettschneider, Alli, ETS, abrettschneider@ets.org
Briggs, Derek, University of Colorado - Boulder, derek.briggs@colorado.edu
Broer, Markus, American Institutes for Research, markus.broer@gmail.com
Brookhart, Susan, Duquesne University, suebrookhart@gmail.com
Brown, Emily, University of North Texas, emilybrown8@my.unt.edu
Brown, Julianna, KnowledgeWorks, brownjc@knowledgeworks.org
Brown, Rachel, The Pennsylvania State University, reb37@psu.edu
Brown, Richard, West Coast Analytics, rich@westcoastanalytics.com
Brown, Terran, Educational Testing Service, txbrown@ets.org
Browy, Madeline, University of Florida, browym@ufl.edu
Brückner, Sebastian, -, brueckner@uni-mainz.de
Brunetti, Matthew, WestEd, mbrunet@wested.org
Brunskill, Emma, Stanford University, ebrun@cs.stanford.edu
Bryer, Jason, City University of New York, jason@bryer.org
Buckendahl, Chad, ACS Ventures, LLC, cbuckendahl@acsventures.com
Buckley, Jack, Roblox, jbuckley@roblox.com
Bueso, Leah, University of Illinois Springfield, lbues@uis.edu
Bufferfuss, Reese, HumRRO, rbutterfuss@humrro.org
Bulut, Okan, University of Alberta, bulut@ualberta.ca
Burkhardt, Amy, Cambium Assessment, burkhardt.amy@gmail.com
Burstein, Jill, Duolingo, jill@duolingo.com
Buzick, Heather, ETS, HBuzick@ets.org
Cai, Li, University of California - Los Angeles, lcai@ucla.edu
Cai, Liuhan, Accenture, sophie.cai@accenture.com
Cai, Yan, Jiangxi Normal University, cy1979123@aliyun.com
Cain, Jennifer, Minnesota Department of Education, jennifer.cain@state.mn.us
Calderón, John Alexander, Instituto Colombiano para la Evaluación de la Educación - ICFES, jcalderon@icfes.gov.co
Caliço, Tiago, Organization for Economic Cooperation and Development, tcalico@proton.me
Camara, Wayne, Law school admissions council, waynecamara@gmail.com
Camargo Salamanca, Sandra Liliana, University of Illinois Urbana-Champaign, sandracs@illinois.edu
Campbell, Ian, Cambium Assessment, ian.campbell@cambiumassessment.com
Can Aran, Ozge, Hacettepe University, ozgecann@hacettepe.edu.tr
Cancado, Luciana, Curriculum Associates, lcancado@cainc.com
Cao, Yichong, The University of Iowa, yichong-cao@uiowa.edu
Kevin Cappaert, Curriculum Associates, KCappaert@cainc.com
Cardwell, Ramsey, Duolingo, ramsey@duolingo.com
Carpenter, Blaire, Florida State University, Bnk11@fsu.edu
Carrasco, Diego, Centro de Medición MIDE UC, Pontificia Universidad Católica de Chile, dacarras@uc.cl
Casabianca, Jodi, ETS, jodi.casabianca@gmail.com
Castillo, Carolina, Pontificia Universidad Católica de Chile, carolina.castillo@uc.cl
Cayumán, Carlos, DEMRE - Universidad de Chile, ccayuman@uchile.cl
Celic, Vuk, UNLV, vuk.celic@unlv.edu
Cervantes, Víctor H, University of Illinois Urbana-Champaign, victorhc@illinois.edu
Chajewski, Michael, Pearson, michael.chajewski@pearson.com
Chakraborty, Roti, Georgia State University, American Institutes for Research, rchakraborty3@student.gsu.edu
Chalhoub-Deville, Micheline, University of North Carolina Greenboro, chalhoub-deville@uncg.edu
Cham, Heining, Fordham University, hcham@fordham.edu
Chang, Chi, Michigan State University, chisq@msu.edu
Chang, Hua-Hua, Purdue University, chang606@purdue.edu
Chang, Joshua, josh@mederrata.com
Chang, Wanchen, Cambium Assessment, claire.chang@cambiumassessment.com

Chao, Hsiu-Yi, Soochow University, psyhyc@scu.edu.tw
Chapin, Grey, University of Florida, greychapin@ufl.edu
Chavez, Carlos, chave143@umn.edu
Chawla, Kamal, University of Maine, kamal.chawla@maine.edu
chen, fangbin, 1098715863@qq.com
Chen, Binhui, bc928@georgetown.edu
Chen, Chi-Chen, National Academy for Educational Research, runrice@hotmail.com
Chen, Hong, University of Iowa, hchen102@uiowa.edu
Chen, Hui-Fang, Department of Social and Behavioural Sciences, City University of Hong Kong, g8932006@gmail.com
Chen, Jianshen, College Board, cachen@collegeboard.org
Chen, Jing, jingchen2010@gmail.com
Chen, Jinsong, The University of Hong Kong, jinsong.chen@live.com
Chen, Jyun-Hong, National Cheng Kung University, psyjhc@gs.ncku.edu.tw
Chen, Qipeng, University of Alabama, qipengchen_psy@outlook.com
Chen, Siyuan, The Ohio State University, chen.8358@osu.edu
Chen, Tingting, The University of Iowa, tingting-chen@uiowa.edu
Chen, Xiuhan, University of Missouri, xcthc@missouri.edu
Chen, Ying, University of Illinois System - IWERC, ychen406@uillinois.edu
Chen, Yinghan, University of Nevada, Reno, yinghanc@unr.edu
Chen, Yunxiao, London School of Economics and Political Science, Y.Chen186@lse.ac.uk
Chen, Zhijun, Department of Education, University of Bath, zc583@bath.ac.uk
Cheng, Britte, Menlo Education Research, bcheng@menloedu.org
Cheng, Ying, University of Notre Dame, ycheng4@nd.edu
Chiu, Chia-Yi, Teachers College, Columbia University, cc5010@tc.columbia.edu
Cho, Youngmi, Riverside Insights, youngmi.cho@riversideinsights.com
Cho, YoungWoo, ACT, youngwoo.cho@act.org
Choe, Edison, Renaissance, edison.choe@renaissance.com
Choi, Hye-Jeong, HumRRO, hchoi@humrro.org
Choi, Ikkyu, ETS, ichoi001@ets.org
Choi, Jean, Curriculum Associates, Jchoi@cainc.com
Choi, Jinah, Edmentum, Jinah.Choi@edmentum.com
Choi, Jinnie, Savvas Learning Company, jinnie.choi@gmail.com
Choi, Mihye, Curriculum Associates, mchoi@cainc.com
Choi, Seung, The University of Texas at Austin, schoi@austin.utexas.edu
Christensen, Laurene, WIDA at the University of Wisconsin-Madison, llchristens2@wisc.edu
Christensen, Wendy, University of Colorado School of Medicine, wchriste@gmail.com
Christopherson, Sara, Wisconsin Center for Education Products and Services (WCEPS), sara@wceps.org
Chu, Man-Wai, University of Calgary, manwai.chu@ucalgary.ca
Church, Jacqueline, Duolingo, jacqueline@duolingo.com
Circi, Ruhan, American Institutes for Research, rcirci@air.org
Cisterna, Dante, ETS, dcisterna@ets.org
Cizek, Gregory, University of North Carolina Chapel Hill, cizek@unc.edu
Clark, Amy, ATLAS: University of Kansas, akclark@ku.edu
Clausen, Mary Kate, Cognia, marykate.clauson@cognia.org
Clauser, Brian, beclauser@gmail.com
Clauser, Jerome, American Board of Internal Medicine, jclauser@abim.org
Clin Deffarges, Alexandre, Institut für Bildungsevaluation Zürich AG; ETH Zurich, aclindeffarg@student.ethz.ch
Close, Catherine, cathynclose@gmail.com
Coates, Adam, University of Bath, ac3615@bath.ac.uk
Cohen, Allan, University of Georgia, acohen946@gmail.com
Cole, Veronica, Wake Forest University, colev@wfu.edu
Colvin, Kimberly, University at Albany, SUNY, kcolvin@albany.edu
Confrey, Jere, North Carolina State University, Department of Science, Technology, Engineering, and Mathematics Education, jconfre@ncsu.edu

Cook, Linda, ETS, lcook@ets.org
Copeland, Nicholas, University of Washington, nrcopela@uw.edu
Copur-Gencturk, Yasemin, University of Southern California, copurgen@usc.edu
Corliss, Mary-Alice, Cognia, mary-alice.corliss@cognia.org
Corter, James, Teachers College, Columbia University, jec34@tc.columbia.edu
Cortes, Juan, DEMRE - Universidad de Chile, juancortes@uchile.cl
Costello, De'Ani, Howard University, de'ani.costello@bison.howard.edu
Courey, Karyssa, Rice University, kac23@rice.edu
Crabtree, Ashleigh, Curriculum Associates, arcrabtree@hotmail.com
Crespo Cruz, Eduardo J, University of Massachusetts, Amherst, ecrespocruz@umass.edu
Crewe, Tara, Denison Montessori, tara_crewe@dpsk12.net
Croft, Michelle, Bellwether, michelle.croft@bellwether.org
Cruce, Ty, ACT, ty.cruce@act.org
Cruz, Alvaro, American Institutes for Research, alcruz@air.org
Cruz, Sandra, DEMRE - Universidad de Chile, spcruz@gwu.edu
Cui, Mengyao, Cambium Assessment, Inc., Mengyao.Cui@cambiumassessment.com
Cui, Weiwei, wcui@collegeboard.org
Cui, Wenju, ETS, wcui@ets.org
Cui, Ying, University of Alberta, yc@ualberta.ca
Cui, Zhongmin, CFA Institute, zhongmin.cui@cfainstitute.org
Culpepper, Steven, University of Illinois at Urbana-Champaign, sculpepp@illinois.edu
Custer, Michael, Riverside Insights, michael.custer@riversideinsights.com
D'Brot, Juan, Center for Assessment, jdbrot@nciea.org
D'sa, Nikhit, University of Notre Dame, ndsa@nd.edu
Dadey, Nathan, The National Center for the Improvement of Educational Assessment, ndadey@nciea.org
Dadzie, Justice, The University of Alabama, jdadzie@crimson.ua.edu
Dai, Shenghai, Washington State University, s.dai@wsu.edu
Dailey, David, David@DaileyDataGroup.com
Daisher, Ted, t3daisher@gmail.com
Dallas, Andrew, drewd@nccpa.net
Danielson, Robert, Washington State University - Spokane, robert.danielson@wsu.edu
Daoust, Carolyn, University of Kansas, cjdaoust@gmail.com
Davidson, Anne, EdMetric, anne.davidson@edmetric.com
Davis, Laurie, Curriculum Associates, laurie@davistx.com
Davis-Becker, Susan, ACS Ventures, LLC, sdavisbecker@acsventures.com
de la Torre, Jimmy, The University of Hong Kong, j.delatorre@hku.hk
De Lisle, Jerome, University of the West Indies - St. Augustine, jeromedelisle@yahoo.com
de los Reyes, Andy, adlr@umd.edu
Deane, Paul, ETS, pdeane@ets.org
Debelak, Rudolf, University of Zurich, rudolf.debelak@psychologie.uzh.ch
Debnam, Candace, School-to-School International, cdebnam@sts-international.org
DeCarlo, Lawrence, Teachers College, Columbia University, decarlo@tc.edu
Delaney, Nora, ndelaney@fas.harvard.edu
Demars, Christine, James Madison University, demarsce@jmu.edu
Demir, Cihan, Washington State University, cihan.demir@wsu.edu
Denbleyker, John, jdenblge@yahoo.com
Deng, Jiayi, Human Resources Research Organization, jiayideng0726@gmail.com
Deng, Nina, Kaplan North America, nina.deng@kaplan.com
DePascale, Charles, Psychometric Confections, LLC, charliedepascale@gmail.com
Deters, Lauren, lauren.fluegge@gmail.com
Di Digiacomo, Daniela, University of Kentucky, daniela.digiacomo@uky.edu
Diao, Hongyu, denisediao@gmail.com
Dicerbo, Kristen, kdicerbo@cox.net
Dickinson, Emily, HumRRO, edickinson@humrro.org

Diegel, Mary Helen, Michigan Department of Education, DiegelM@michigan.gov
Ding, Jiyuan, dingjiyuan1210@163.com
DiRenzo, Raegan, University of Florida, raegandirenzo@ufl.edu
Disale, Ranjitsinh, HGSE, rdisale@gse.harvard.edu
DiStefano, Christine, University of South Carolina, distefan@mailbox.sc.edu
Dobria, Lidia, Wilbur Wright College, LDobria@ccc.edu
Domingue, Benjamin, Stanford University, ben.domingue@gmail.com
Dong, Yixiao, University of Denver, yixiao.dong@du.edu
Donoghue, John, ets, jdonoghue@ets.org
Doran, Harold, HumRRO, hdoran@humrro.org
Douglas, Kerrie, Purdue University, douglask@purdue.edu
Drake, Philine, DIPF, p.drake@dipf.de
Draney, Karen, kdraney@berkeley.edu
Dreyer, Stacy, ACT Education Corp, stacy.dreyer@act.org
Drost, Bryan, Rocky River City School District, drost.bryan@rrcs.org
Du, Ying, American Board of Pediatrics, ydu@abpeds.org
Duan, Qizhou, qduan@nd.edu
Dumas, Denis, University of Georgia, Denis.Dumas@uga.edu
Dunbar, Stephen, University of Iowa, steve-dunbar@uiowa.edu
Dunn, Jennifer, Pearson, jenn.dunn@pearson.com
Duran, Lillian, University of Oregon, lduran@uoregon.edu
Durmus, Aybuke, TED University, aybuke.durmus@tedu.edu.tr
Dwyer, Andrew, American Board of Pediatrics, adwyer@abpeds.org
Dym, Abigail, Providence College, abigaild@ucr.edu
E. Tran, Jasmine, Stanford University, jasetran@stanford.edu
Eagan, Brendan, University of Wisconsin Madison, beagan@wisc.edu
Eckerly, Carol, ETS, caroleckerly@gmail.com
Eckes, Thomas, University of Bochum, eckes@gast.de
Edeh, Ejike, ejedeh@uark.edu
Edi, Daniel, NC Department of Public Instruction, edidatasolutioneds@gmail.com
Edwards, Kelly, American Board of Pediatrics, kedwards@abpeds.org
Effatpanah Hesari, Farshad, Technische Universität Dortmund, farshad.effatpanah@tu-dortmund.de
Egan, Karla, EdMetric, karla.egan@edmetric.com
Eggers-Robertson, Lisa, lisa.eggers@pearson.com
Embretson, Susan, Georgia Institute of Technology, susan.embretson@psych.gatech.edu
Emery, Matt, Roblox, memery@roblox.com
Emler, Trina, University of Kansas, AAI, CLARA, trina.emler@ku.edu
Engelhard, George, University of Georgia, gengelh@uga.edu
Englert, Kerry, Seneca Consulting, LLC, kenglert@senecaconsulting.org
Ercikan, Kadriye, ETS, kercikan@ets.org
Erickson, Lori, lori.erickson@spps.org
Escribano, Rosario, Pontificia Universidad Catolica de Chile, Facultad de Educación, rosario.escribano@uc.cl
Espinosa Aguirre, María Jesús, Universidad Diego Portales, Facultad de Educación, maria.espinosaa@mail.udp.cl
Estrada, Samantha, The University of Texas at Tyler, estr2525@gmail.com
Evans, Carla, Center for Assessment, cevans@nciea.org
Everson, Howard, howard.everson@gmail.com
Eze, Jesse, University of Delaware, ezejesse@udel.edu
Ezugwu, Ifesinachi, The University of Alabama, ijezugwu@crimson.ua.edu
Fan, Fen, National Board of Medical Examiners, ffan@nbme.org
Fan, Meng, HUMRRO, mfan@humrro.org
Fan, Ziyu, The University of Texas at Austin, ziyufan1219@outlook.com
Fancsali, Steven, Carnegie Learning, sfancsali@carnegielearning.com
Fang, Luyang, University of Georgia, Luyang.Fang@uga.edu
Fauss, Michael, ETS, mfauss@ets.org

Fay, Derek, Pearson, derek.m.fay@gmail.com
Federiakin, Denis, Johannes Gutenberg University of Mainz; Goethe University of Frankfurt, denis.federiakin@gmail.com
Feinberg, Richard, NBME, RFeinberg@nbme.org
Feng, Jing, College Board, jfeng@collegeboard.org
Feng, Mingyu, WestEd, mfeng@wested.org
Feng, Xiaoying, American Institutes for Research, ctr_xfeng@air.org
Feng, Ying, ETS, yfeng@ets.org
Feng, Zechu, The University of Hong Kong, jason521@connect.hku.hk
Ferrara, Steve, HumRRO, sferrara1951@gmail.com
Ferris, Julia, City University of New York, julia.ferris03@spsmail.cuny.edu
Fickes, Natalie, Denison Montessori, natalie_fickes@dpsk12.net
Figueroa Millares, Javiera, Universidad Alberto Hurtado, Facultad de Educación, jafigueroa@uahurtado.cl
Filonczuk, Audrey, afiloncz@nd.edu
Fina, Anthony, University of Iowa, anthony-fina@uiowa.edu
Finch, W. Holmes, Ball State University, whfinch@bsu.edu
Fincher, Melissa, edCount, LLC, mfincher@edcount.com
Fink, Aron, Johann Wolfgang Goethe University, a.fink@psych.uni-frankfurt.de
Finney, Sara, James Madison University, finneysj@jmu.edu
Firoozi, Tahereh, University of Alberta, tahereh@ualberta.ca
Fisk, Charles, NBOME, fisk@umd.edu
Fitzsimmons, Ellen, University of Missouri, eafbym@mail.missouri.edu
Flor, Michael, ETS, mflor@ets.org
Fong, Michael, Southern Methodist University, fongm@mail.smu.edu
Forsyth, Carol, ETS, cforsyth@ets.org
Forte, Ellen, edCount, LLC, eforte@edcount.com
Foster, Chris, Caveon, Chris.Foster@caveon.com
Foster, David, david.foster@caveon.com
Foster, Paul, Southern Methodist University, pdfoster@smu.edu
Fraser, Antonio, City University of New York, tony.fraser@gmail.com
Fremer, John, Caveon Test Security, john.fremer@caveon.com
French, Brian, Washington State University, frenchb@wsu.edu
Freund, Dave, ETS, dfreund@ets.org
Freund, Rebecca, Pearson, rebecca.freund@pearson.com
Frey, Andreas, Goethe University Frankfurt, frey@psych.uni-frankfurt.de
Fu, Jianbin, ETS, jfu@ets.org
Fu, Yanyan, Graduate Management Admission Council, frankyanyan@gmail.com
Fuentes-Jimenez, Mia, miafuen@stanford.edu
Fulmore, Julia, University of Dallas, jfulmore@udallas.edu
Furgol Castellano, Katherine, ETS, KEcastellano@ets.org
Gabler, Jennifer, Johannes Gutenberg University Mainz, jenny.fischer@uni-mainz.de
Gallagher, Melissa, U.S. Math Recovery Council, mgallagher@mathrecovery.org
Gan, Kaiyu, Beijing Normal University, bnugankaiyu@163.com
Gandara, Fernanda, Room to Read, fernanda.gandara@roomtoread.org
Gandhi, Kanishk, Stanford University, kanishk.gandhi@stanford.edu
Gane, Brian, bgane@ku.edu
Gao, Yizhu, University of Georgia, Yizhu.Gao@uga.edu
Ge, Yuan, College Board, yge@collegeboard.org
Gearin, Brian, University of Oregon, Brian Gearin, bgearin@uoregon.edu
Geisinger, Kurt, kgeisinger2@unl.edu
Gerker, Heather, University of Kansas, hgerker@ku.edu
Ghonsooly, Behzad, Ferdowsi University of Mashhad, ghonsooly@um.ac.ir
Gianopulos, Garron, Cambium Assessment, garron.gianopulos@cambiumassessment.com
Gierl, Mark, University of Alberta, mgierl@ualberta.ca
Giers, Benjamin, University of Calgary, ben.giers@ucalgary.ca

Giers, David, University of Calgary, giersd.dg@gmail.com
Gilbert, Joshua, Joshua_gilbert@gse.harvard.edu
Gill, Matthew, South Dakota Department of Education, matthew.gill@state.sd.us
Gilmore, Jack, James Madison University, gilmorjr@jmu.edu
Gitiria, Lucy, UMass Amherst, gitt610@gmail.com
Gkolia, Aikaterini, University of Nicosia, gkolia.a@unic.ac.cy
Glore, Christy, ATLAS, University of Kansas, christyglore@ku.edu
Gochyyev, Perman, University of California, Berkeley, perman@berkeley.edu
Godoy, María Inés, DEMRE - Universidad de Chile, maria.godoy.a@uchile.cl
Golden, Richard, golden@utdallas.edu
Gong, Brian, Center for Assessment, bgong@nciea.org
Gonzalez, Mario, Cambium Assessment, mario.gonzalez@cambiumassessment.com
Gooch, Reginald, ETS, rmgooch@ets.org
Goodman, Joshua, National Commission on Certification of Physician Assistants, joshuag@nccpa.net
Goodman, Noah, Stanford University, ngoodman@stanford.edu
Goodwin, Sarah, Duolingo, sarahg@duolingo.com
Gopalakrishnan, Madhumitha, New York University, mg7584@nyu.edu
Gordon, E. Wyatt, Pearson, wyatt.gordon@pearson.com
Gorgulu, Rumeysa, Purdue University, rgorgul@purdue.edu
Gorgun, Guher, University of Alberta, gorgun@ualberta.ca
Gorin, Joanna, ACT Education Corp., joanna.gorin@act.org
Gorney, Kylie, Michigan State University, kgorney@msu.edu
Gorno Tempini, Marilu, University of California, San Francisco, marialuisa.gornotempini@ucsf.edu
Gorum Bing, Crystal, National Board of Medical Examiners, cbing@nbme.org
Gotzmann, Andrea, Medical Council of Canada, agotzmann@mcc.ca
Goulart, Sarah, UC Berekeley, sarahgoulart@berkeley.edu
Graf, Edith, ETS, agraf@ets.org
Gregg, Nikole, National Commission on Certification of Physician Assistants, ngregg11@gmail.com
Greiert, Brittany, Aurora Public Schools, bsgreiert@aurorak12.org
Grekov, Paulina, University of Wisconsin-Madison, pgrekov@wisc.edu
Gribben, Monica, HumRRO, MGribben@humrro.org
Griger, Cassondra, The University of Iowa, cassondragriger@gmail.com
Grochowalski, Joe, College Board, jgrochowalski@collegeboard.org
Gui, Yi, The University of Iowa, yi-gui@uiowa.edu
Guo, Enruo, ETS, eguo@ets.org
Guo, Hongwen, ETS, hguo@ets.org
Guo, Shuchen, University of Georgia, Shuchen.Guo@uga.edu
Guo, Wenjing, Pearson, wenjing.guo@pearson.com
Guo, Yage, Center for Applied Linguistics, yguo@cal.org
Guo, Yanhui, University of Illinois Springfield, yguo56@uis.edu
Gupta, Vishal, Indian Institute of Management Ahmedabad, vishal@iima.ac.in
Gushta, Matthew, MetaMetrics, Inc., mgushta@lexile.com
Haab, Sergio, The University of Iowa, shaab@uiowa.edu
Haber, Nick, Stanford University, nhaber@stanford.edu
Habing, Brian, University of South Carolina, habing@stat.sc.edu
Hadi, Syed Abdul, shadi@wisc.edu
Halai, Anjum, Aga Khan University, anjum.halai@aku.edu
Halpin, Peter, UNC-Chapel Hill, peter.halpin@unc.edu
Ham, Dae Woong, University of Michigan, daewoongham@g.harvard.edu
Ham, Eun Hye, Kongju National University, thanks02@gmail.com
Hamdani, Maria, The Center for Measurement Justice, mhamdani@measurementjustice.org
Hamilton, Laura, American Institutes for Research, lhamilton@air.org
Hammami, Sarah, shammami2@huskers.unl.edu
Han, Kyung (Chris), GMAC, truetheta@gmail.com

Han, Suhwa, Cambium Asessment, suhwa@utexas.edu
Han, Youngjin, University of Maryland, younghan@umd.edu
Han, Yuting, Beijing Language and Culture University, hanyuting716@gmail.com
Hao, Jiangang, ETS, jhao@ets.org
Hao, Kevin, Duolingo, kevinhao1234@gmail.com
Harik, Polina, NBME, pharik@nbme.org
Harra, Kjorte, University of Wisconsin-Madison, kjorteh@gmail.com
Harrold, Archer, University of Michigan, aharrold@umich.edu
Harrold, Brian, University of South Carolina, bharrold@email.sc.edu
Hartig, Johannes, DIPF, j.hartig@dipf.de
Hass, Christopher, Rutgers University, chris.hass@shaw.ca
Hatch, Amber, University of Florida, hatcha@ufl.edu
Hauenstein, Clifford, Johns Hopkins School of Medicine, chauens2@jh.edu
Haviland, Sara, ETS, shaviland@ets.org
He, Joy, Stanford University, heyueya@stanford.edu
He, Qiwei, Georgetown University, qiwei.he@georgetown.edu
He, Surina, surina1@ualberta.ca
He, Yi, Edmentum, yi.he@edmentum.com
Headley, Mollie, Curriculum Associates, mheadley@cainc.com
Hebert, Andrea, Cognia, andreahebert66@gmail.com
Heckler, Andrew, The Ohio State University, heckler.6@osu.edu
Hefty, Tyler, Renaissance, tyler.hefty@renaissance.com
Hellman, Scott, Pearson, scott.hellman@pearson.com
Hembry, Tracey, Pearson, tracey.hembry@pearson.com
Hemphill, Cadelle, American Institutes for Research, chemphill@air.org
Henderek, Jerusha, NBME, JHenderek@nbme.org
Hendrickson, Amy, The College Board, ahendrickson@collegeboard.org
Henríquez, Catalina, University of Colorado - Boulder, catalina.henriquez@colorado.edu
Henson, Robert, University of North Carolina Greensboro, rahenson@uncg.edu
Hernandez, Jasmin, UC Irvine, jasmih10@uci.edu
Hernández, Diley, Georgia Institute of Technology, diley.hernandez@gatech.edu
Herr, Riley, James Madison University, herrrk@jmu.edu
Herrada, Maximiliano, DEMRE - Universidad de Chile, maximiliano.herrada@uchile.cl
Hicks, Juanita, juanita.hicks.11@gmail.com
Hicks, Lloyd, Westat, LloydHicks@westat.com
Himelfarb, Igor, ihimelfarb@nbce.org
Hinkle, Andrew, National Center on Educational Outcomes, University of Minnesota, ahinkle@umn.edu
Ho, Andrew, Harvard University, Andrew_Ho@gse.harvard.edu
Ho, TsungHan, ETS, tho@ets.org
Hodes, Paul, hodes@uni-mainz.de
Hodge, Kari, Ascend Learning, kari.hodge@ascendlearning.com
Hodge-Clark, Kristen, Georgetown University School of Continuing Studies, kh866@georgetown.edu
Hoeve, Karen, American Board of Pediatrics, khoeve@abpeds.org
Hoffman, Alexander, AleDev Research & Consulting, ahoffman@AleDev.com
Hojnoski, Robin, Lehigh University, roh206@lehigh.edu
Holder, Eli, eli@3iap.co
Holliday, Keith, Curriculum Associates, kholliday@cainc.com
Hong, Hyeri, California State University, Fresno, hyerihong@mail.fresnostate.edu
Hong, IkHyeon, Inha University, hih84562@gmail.com
Hong, Seong Eun, Center for Applied Linguistics, shong@cal.org
Hooks-Singletary, Brianna, University of North Carolina Greenboro, briannahooks41@gmail.com
Hopfenbeck, Therese, University of Melbourne, therese.hopfenbeck@unimelb.edu.au
Horton, Omaya, University of Massachusetts Amherst, ohorton@umass.edu
Hoskins, Jessica, jess.es.hoskins@gmail.com

Hou, Tiantian, Beijing Normal University, deriase@163.com
Houang, Richard, Michigan State University, houang@msu.edu
Houriet, Andrew, NBME, ahouriet@nbme.org
Hove, Jennifer, Education Quality and Accountability Office, jennifer.hove@eqao.com
Hsiao, I-Chun Vera, University of Iowa, i-chun-hsiao@uiowa.edu
Hsieh, Hsiaolin, Stanford University, hsiaolin@stanford.edu
Hu, Sara, Education Analytics, shu@edanalytics.org
Hu, Yian, yh813@georgetown.edu
Hua, Cheng, University of Montevallo, chua@montevallo.edu
Hua, Haowei, The Culver Academies, 654053@culver.org
Huang, Anne, Scarsdale High School, New York, annexahuang7@gmail.com
Huang, Chi-Yu, ACT, chiyu.huang@act.org
Huang, Jing, jing.huang1222@gmail.com
Huang, Mingya, University of Wisconsin -- Madison, mhuang233@wisc.edu
Huang, Qi, University of Wisconsin - Madison, qhuang85@wisc.edu
Huang, Sijia, Indiana University Bloomington, sijhuang@iu.edu
Huang, Wei, Universtiy of Alabama, College of Education, whuang20@crimson.ua.edu
Huang, Yingshi, University of California - Los Angeles, yingshi@ucla.edu
Huang, Yuan, New York University, yh2741@nyu.edu
Huang, Yue, Measurement Incorporated, yueh@udel.edu
Huff, Kristen, Curriculum Associates, khuff@cainc.com
Huggins-Manley, Corinne, amanley@coe.ufl.edu
Hughes, Sarah, sarah.hughes1@pearson.com
Hunt, Jessica, North Carolina State University, jhunt5@ncsu.edu
Hwang, Heungsun, McGill University, heungsun.hwang@mcgill.ca
Hwang, Kyu-Ryng, New Hampshire Department of Education, kyu-ryung.hwang@doe.nh.gov
Hwang, Yaurie, The Learning Partnership, yaurie@thelearningpartnership.net
Iaccarino, Stephanie, NBME, siaccarino@nbme.org
Ignacio, Angelie, University of Toronto, angelie.ignacio@utoronto.ca
Ihlenfeldt, Samuel, University of Minnesota, ihlen010@umn.edu
Im, Sukkeun, Pearson, sukkeun@gmail.com
Immanuel, Aria, University of Massachusetts Amherst, aimmanuel@umass.edu
Immekus, Jason, jcimme01@louisville.edu
Ing, Marsha, University of California, Riverside, marshai@ucr.edu
Ingrisone, James, Pearson VUE, james.ingrisone@pearson.com
Ingrisone, Soo, HumRRO, singrisone@gmail.com
Isbell, Daniel, University of Hawaii at Manoa, disbell@hawaii.edu
Ivanova, Militsa, University of Cyprus, militsagi@yahoo.com
Iyer, Balaji, University of Florida, balaji.iyer@ufl.edu
Jackson Haynes, Janine, jajac64@morgan.edu
Jahan, Israt, University of Calgary, israt.jahan1@ucalgary.ca
Jain, Vipul, Michigan State University, jainvip1@msu.edu
James, Miriam, International Centre for Educational Evaluation, Institute of Education, University of Ibadan, miriamjames700@gmail.com
Jang, Eunice, University of Toronto, eun.jang@utoronto.ca
Jarr, Karoline, karolinejarr@hotmail.com
Jeffries, Jay, University of Nebraska-Lincoln, jayjeffries13@huskers.unl.edu
Jensen, Bryant, Brigham Young University, bryant_jensen@byu.edu
Jeon, Minjeong, UCLA, mjjeon@ucla.edu
Jewsbury, Paul, ETS, pjewsbury@ets.org
Ji, Feng, University of Toronto, f.ji@utoronto.ca
Ji, Jeongmin, University of Iowa, jeongmin-ji@uiowa.edu
Ji, Xuejun, x.ryan.ji@gmail.com
Jia, Ao, Jiangxi Normal University, 17799379956@163.com

Jia, Helena, Educational Testing Service, YJia@ets.org
Jiang, Yang, ETS, yjiang002@ets.org
Jiang, Yanlin, NCCPA, yanlinj@nccpa.net
Jiang, Yanming, HMH, yanming.jiang@hmhco.com
Jiao, Hong, University of Maryland, hjiao@umd.edu
Jimenez, Auburn, University of Kansas, auburn.jimenez@ku.edu
Jimenez, Daniela, DEMRE - Universidad de Chile, da.jimenez@uchile.cl
Jimenez-Silva, Margarita, University of California - Davis, mjimenezsilva@ucdavis.edu
Jin, Kuan-Yu, Hong Kong Examinations and Assessment Authority, kyjin@hkeaa.edu.hk
Jin, Rong, NBOME, rongjin2012@gmail.com
Jin, Ying, Association of American Medical Colleges, yjin@aamc.org
Johanningmeier, Scott, ACT Education Corp, Scott.Johanningmeier@act.org
Johnson, Evelyn, Riverside Insights, ejohnson@apertureed.com
Johnson, Matthew, University of South Florida, matthewjohnson@usf.edu
Johnson, Matthew, ETS Research Institute, msjohnson@ets.org
Johnson, Odis, Johns Hopkins University; ETS Research Institute, ojohnson@ets.org
Jones, Andrew, American Board of Surgery, ajones@absurgery.org
Joo, Joanne, Southern Methodist University, jjoo@smu.edu
Joo, Sean, University of Kansas, sjoo@ku.edu
Joshi, Malt, Texas A&M University, mjoshi@tamu.edu
Ju, Unhee, Cognia, unhee.ju@cognia.org
Julien, Nisha Acharya, Center for Measurement Justice, nacharya@measurementjustice.org
Julian, Marc, DRC, mjulian@datarecognitioncorp.com
Jung, Ae Kyong, University of Iowa, aekyong-jung@uiowa.edu
Jung, Juyoung, University of Iowa, juyoung-jung@uiowa.edu
Kılıç, Abdullah, Trakya University, abdullahfarukkilic@gmail.com
Kaçak, Tugay, Trakya University, kacaktugay@gmail.com
Kahne, Joseph, University of California, Riverside, jkahne@ucr.edu
Kamata, Akihito, akamata@smu.edu
Kane, Michael, mkane@ets.org
Kang, Hyeon-Ah, University of Texas Austin, hkang@austin.utexas.edu
Kang, Youngsoon, Amira+Istation, dr.yskang22@gmail.com
Kannan, Priya, WestEd, pkannan@wested.org
Kanopka, Klint, New York University, klint.kanopka@nyu.edu
Kantor, Cayce, University of Denver, cayce.kantor@du.edu
Kaplan, David, University of Wisconsin - Madison, david.kaplan@wisc.edu
Kapoor, Shalini, shalinikapoor.ia@gmail.com
Kaptur, Bradley, HSHS Saint John's Hospital, bdkaptur@gmail.com
Kaptur, Danielle (Dandan), Pearson, dandan.c.chen@gmail.com
Kara, Yusuf, University of Miami, ykara@miami.edu
Karamese, Hacer, WIDA at the University of Wisconsin-Madison, karamese@wisc.edu
Karelitz, Tzur, National Institute for Testing & Evaluation (NITE), tzur@nite.org.il
Karimi Malekabadi, Farzan, karimima@usc.edu
Karvonen, Meagan, karvonen@ku.edu
Kary, David, Law School Admission Council, dkary@lsac.org
Kehat, Gitit, Cambium Assessment, gitit.kehat@cambiumassessment.com
Kehinde, Olasunkanmi, Elizabeth City State University, ojkehinde@ecsu.edu
Kelleigh, Willow, The Learning Partnership, wkelleigh@thelearningpartnership.net
Keller, Brian, University of Missouri, btkeller@missouri.edu
Keller, Lisa, University of Massachusetts, Amherst, lkeller@educ.umass.edu
Keller, Rob, HumRRO, rob@robkeller.com
Keng, Leslie, National Board of Medical Examiners, lesliekeng@gmail.com
Kennedy, Patrick, University of Oregon, ppaine@uoregon.edu
Kern, Justin, University of Illinois at Urbana-Champaign, kern4@illinois.edu

Kerzabi, Emily, ETS, ekerzabi@ets.org
Ketterlin Geller, Leanne, Southern Methodist University, lkgeller@mail.smu.edu
Keum, Eunhee, CRESST, keum@cresst.org
Kevelson, Marisol, ETS, mkevelson@ets.org
Khurana, Havisha, University of Oregon, havishak@uoregon.edu
Kidd, David, Harvard University, david_kidd@gse.harvard.edu
Killion, Sarah, Aurora Public Schools, skkillion@aurorak12.org
Kim, Doyoung, Korea Institute for Curriculum and Evaluation, dypark@kice.re.kr
Kim, Edward, Bentley University, ekim@bentley.edu
Kim, Eunbee, Georgia Institute of Technology, eunbee.kim@gatech.edu
Kim, Hahyeong, University of Illinois at Urbana-Champaign, hk33@illinois.edu
Kim, Hyunah, Education Quality and Accountability Office, hyunah.kim@eqao.com
Kim, Hyunjoo, hyunjoo5@illinois.edu
Kim, Jieun, University of Hawaii at Manoa, jieun247@hawaii.edu
Kim, JongPil, Riverside Insights, jp.kim@riversideinsights.com
Kim, Kyung Yong, University of North Carolina Greensboro, k_kim9@uncg.edu
Kim, Meereem, Korea Institute for Curriculum and Evaluation, meereemkim@gmail.com
Kim, Minjung, mjkim.jayhawk@gmail.com
Kim, Nana, University of Minnesota, nkim530@umn.edu
Kim, Seongeun, s_kim45@uncg.edu
Kim, Stella, University of North Carolina Charlotte, stella-kim@charlotte.edu
Kim, Sungyeun, Incheon National University, syk@inu.ac.kr
Kim, Yejin, dpwls3360@snu.ac.kr
Kim, YJ, yj.kim@wisc.edu
Kim, Young Yee, American Institutes for Research, ykim@air.org
Kim, YoungKoung, The College Board, ykim@collegeboard.org
Kim, Youngwon, Harvard University, youngwonkim.ywk@gmail.com
Kim, Yun-Kyung, UCLA/CRESST, yunkim2729@ucla.edu
King, Emma, American Institutes for Research, eking@air.org
Kinsey, Devon, ETS, DKINSEY001@ets.org
Kingston, Neal, University of Kansas, nkingston@ku.edu
Kintz, Tara, Michigan Assessment Consortium, kintztar@msu.edu
Kirakosian, Antranik, Washington State University, antranik.kirakosian@wsu.edu
Kislyonkova, Alexandra, University of California Merced, akislyonkova@ucmerced.edu
Klein, Trish, ENTK LLC, trishklein44@gmail.com
Koehn, Hans, University of Illinois at Urbana-Champaign, hkoehn@illinois.edu
Kohmer, Anika, Johannes Gutenberg University Mainz, ankohmer@uni-mainz.de
Kokolis, Gerry, ETS, gkokolis@ets.org
Kong, Lingchen, University of Florida, l.kong@ufl.edu
König, Christoph, Goethe University Frankfurt, Germany, koenig@psych.uni-frankfurt.de
Konstantinidou, Evi, University of Nicosia, konstantinidou.ev@unic.ac.cy
Koo, Miryeong, University of Illinois at Urbana-Champaign, koo21@illinois.edu
Koval, Jayma, Georgia Institute of Technology, jayma.koval@ceismc.gatech.edu
Kozikowski, Andrzej, NCCPA, andrzejk@nccpa.net
Kremmel, Benjamin, University of Innsbruck, Benjamin.Kremmel@uibk.ac.at
Krenzke, Tom, Westat, TomKrenzke@westat.com
Krist, Andrew, atkrist@crimson.ua.edu
Kroopnick, Marc, Association of American Medical Colleges, mkroopnick@aamc.org
Krost, Kevin, kevinkrost@vt.edu
Krumm, Andrew, aekrumm@umich.edu
Krupar, Allyson, Save the Children, akrupar@savechildren.org
Kuang, Huan, Florida State University, hkuang2@fsu.edu
Kukea Shultz, Pohai, University of Hawaii - Manoa, pohai@hawaii.edu
Kulczycki, Christy, Cambium Assessment, Inc., christy.kulczycki@cambiumassessment.com

Kuncel, Nathan, kunce001@umn.edu
Kunina-Habenicht, Olga, Technische Universität Dortmund, Olga.Kunina-Habenicht@tu-dortmund.de
Kunz, Ann-Kathrin, Johannes Gutenberg University Mainz, a-k.kunz@uni-mainz.de
Kuo, Tzu-Chun, Kaplan North America, tzuchun.kuo@kaplan.com
Kush, Joseph, James Madison University, kushjm@jmu.edu
Kwako, Alexander, Cambium Assessment, Alexander.Kwako@cambiumassessment.com
Kwon, Muwon, University of Maryland, mwkwon21@umd.edu
Kwon, Sunbeom, University of Illinois, Urbana-Champaign, sunbeom2@illinois.edu
Kyllonen, Patrick, ETS, pkyllonen@ets.org
Kyung, Jungwon, University of Massachusetts, Amherst, jkyung@umass.edu
Ladson-Billings, Gloria, University of Wisconsin-Madison, gjladson@wisc.edu
Laesecke, Anne, USAID, alaesecke@usaid.gov
LaFlair, Geoff, Duolingo, geoff@duolingo.com
Lahoud, Tamlyn, University of Georgia, tamlyn.lahoud@uga.edu
Lai, Hok Chio, University of Southern California, hokchiol@usc.edu
Lai, Viet, Adobe Research, viet.ptit.bn@gmail.com
Laitusis, Cara, Center for Assessment, claitusis@nciea.org
Lakin, Joni, University of Alabama, Jlakin@ua.edu
Lan, Andrew, University of Massachusetts, Amherst, andrewlan@cs.umass.edu
Lane, John, Michigan Assessment Consortium, lanejoh3@msu.edu
Langi, Meredith, mlangi@wested.org
Laurin, Charles, Institute of Education Sciences - US Department of Education, charles.laurin@ed.gov
Lavigne, Cheryl, ETS, clavigne@ets.org
Lazarus, Sheryl, laza0019@umn.edu
Le, Anh Thu, University of Notre Dame, ale3@nd.edu
LeBeau, Brandon, WestEd, blebeau@wested.org
Lederman, Josh, Brandeis University, josh.lederman.edu@gmail.com
Lee, Chansoon (Danielle), dchanslee@gmail.com
Lee, Haeju, hlee@uncg.edu
Lee, Haneul, Inha University, gksmf6770@naver.com
Lee, Hyeryung, hyerlee@uiowa.edu
Lee, Hyunjung, Fordham University, hlee201@fordham.edu
Lee, Jeremy, ETS, jylee001@ets.org
Lee, Jiyoon, Ewha Womans University, jyn.lee@ewha.ac.kr
Lee, Mina, Cambium Assessment, mina.lee@cambiumassessment.com
Lee, Minho, University of California - Los Angeles, leemino72@ucla.edu
Lee, Selene, ETS, selene0525@gmail.com
Lee, Sooyong, WIDA at the University of Wisconsin-Madison, slee2462@wisc.edu
Lee, Sung-Hyuck, GMAC, lee9410@hotmail.com
Lee, Sunhyoung, Ascend Learning/ University of Nebraska-Lincoln, hongasunny@gmail.com
Lee, Won-Chan, University of Iowa, won-chan-lee@uiowa.edu
Lee, Yongsang, Inha University, yong21c@gmail.com
Leeming, Meghan, University of North Carolina Greensboro, mkleemin@uncg.edu
Lefstein, Adam, The Hebrew University of Jerusalem, adam.lefstein@mail.huji.ac.il
Lehman, Blair, Brighter Research, blehman@brighter-research.com
Lekwa, Adam, Rutgers, The State University of New Jersey, adam.lekwa@rutgers.edu
Lembke, Erica, University of Missouri, lembkee@missouri.edu
Leng, Dihao, Boston College, dihaoleng@outlook.com
Lentini, Maria, Pearson, maria.rlentini@pearson.com
Lerch, Ryan, lerch013@g.ucla.edu
LeRoy, Stephanie, leroysk@jmu.edu
Leventhal, Brian, James Madison University, leventbc@jmu.edu
Levi, Effi, National Institute for Testing & Evaluation (NITE), effi@nite.org.il
Lewis, Daniel, Creative Measurement Solutions, dan.lewis@creativemeasurement.com

Lewis, Jennifer, University of Massachusetts Amherst, jlewi0@umass.edu
Lewis, John, Denver Public Schools, john_lewis1@dpsk12.net
Lewis, Regina, Maine Department of Education, regina.lewis@maine.gov
Li, Dongmei, dongmei.li@act.org
Li, Feiming, Zhejiang Normal University, feimingli@zjnu.edu.cn
Li, Grace, WIDA at the University of Wisconsin – Madison, grace.li@wisc.edu
Li, Ji, University of Arkansas, jili@uark.edu
Li, Jian, Beijing Normal University, jianli@bnu.edu.cn
Li, Jiawei, Florida State University, jl14w@fsu.edu
Li, Jie, lijdbc@gmail.com
Li, Jujia, University of Alabama, jli183@crimson.ua.edu
Li, Lanrong, MetaMetrics, Inc., jessicalilr2011@gmail.com
Li, Min, University of Washington, minli@u.washington.edu
Li, Ming, University of Maryland, minglii@umd.edu
Li, Shuhong, Educational Testing Service, sli@ets.org
Li, Shuyuan, Education Analytics, sli@edanalytics.org
Li, Warren, Associate Research Scientist, warren@ld-insights.com
Li, Wenshuo, University of Alberta, wenshuo5@ualberta.ca
Li, Xiao, University of South Florida, xli@usf.edu
Li, Xin, xin.li@act.org
Li, Yifei, lieln@bc.edu
Li, Zhifei, Beijing Normal university, lizhifei@mail.bnu.edu.cn
Liang, Min, National Board of Osteopathic Medical Examiners (NBOME), min-liang-1@uiowa.edu
Liang, Xinya, University of Arkansas, xl014@uark.edu
Liao, Manqian, Duolingo, mancy@duolingo.com
Liao, Xiangyi, University of Wisconsin Madison, xliao36@wisc.edu
Liaw, Yuan-Ling, IEA Hamburg, yuan-ling.liaw@iea-hamburg.de
Lim, Hwanggyu, Inha University, hglim83@gmail.com
Lim, Sangdon, Cambium Assessment, stdevlimit@gmail.com
Lim, Youn Seon, University of Cincinnati, limyo@ucmail.uc.edu
Lin, Pei-Ying, University of Saskatchewan, pei-ying.lin@usask.ca
Lin, Ye, Ascend Learning, yelin.nora@gmail.com
Lindner, Marlit, IWM - Leibniz Institut fŸr Wissensmedien, m.lindner@iwm-tuebingen.de
Ling, Guangming, ETS, gling@ets.org
Lipner, Rebecca, rlipner@abim.org
Litschwartz, Sophie, MDRC, sophie.litschwartz@mdrc.org
Liu, Chunyan, National Board of Medical Examiners, cliu@nbme.org
Liu, Hongyun, Beijing Normal University, hyliu@bnu.edu.cn
Liu, Huan, Riverside Insights, huan.liu@riversideinsights.com
Liu, Jin, The University of Texas at Arlington, jin.liu@uta.edu
Liu, Jinghua, The National Board of Osteopathic Medical Examiners, jinghuawalkerliu@gmail.com
Liu, Joyce Xinle, University of Alberta, xinle4@ualberta.ca
Liu, Junhui, jliu001@ets.org
Liu, Lei, ETS, lliu001@ets.org
Liu, Lucia, Ascend Learning, lucy.xin.liu@gmail.com
Liu, Ou Lydia, ETS, lliu@ets.org
Liu, Qiao, qliu14@charlotte.edu
Liu, Qidi, GlobalFoundries, qidi.liu.photonics@gmail.com
Liu, Ran, Amira Learning, ran.liu@amiralearning.com
Liu, Ren, University of California, Merced, rliu45@ucmerced.edu
Liu, Ruitao, 2U, rliuufl@gmail.com
Liu, Xiaoxiao, University of Alberta, xiaoxia6@ualberta.ca
Liu, Xinyu, jane1023662040@163.com
Liu, Yang, University of Maryland, yliu87@umd.edu

Liu, Yaohui, University of Alabama, yliu73@ua.edu
Liu, Ying, University of Illinois at Urbana-Champaign, yingl7@illinois.edu
Liu, Yunting, University of California, Berkeley, yunting99@berkeley.edu
Lnu, Chandramani, University of Maryland, cm05@umd.edu
Lo, Kai-Ling, Duolingo, kailing@duolingo.com
Lochbaum, Karen, Pearson, Karen.Lochbaum@pearson.com
Lockspeiser, Tai, University of Colorado School of Medicine, tai.lockspeiser@cuanschutz.edu
Lockwood, J.R., Duolingo, jr@duolingo.com
Long, Haiying, University of Kansas, hlong@ku.edu
Long, Yiting, Westat, YitingLong@westat.com
López, Dany, DEMRE - Universidad de Chile, dany.lopez@uchile.cl
Lottridge, Susan, Cambium Assessment, susan.lottridge@cambiumassessment.com
Lou, Yaosheng, University at Albany, SUNY, ylou@albany.edu
Lovelace, Temple, Assessment for Good, AERDF, tlovelace@aerdf.org
Lu, Ru, Educational Testing Service, rlu@ets.org
Lu, Xinyi, University of Michigan, lwlxy@umich.edu
Lu, Yikai, University of Notre Dame, ylu22@nd.edu
Lu, Ying, The College Board, ylu@collegeboard.org
Lu, Yining, yl11897@nyu.edu
Luccioli, Viviana, Georgetown University, vcl16@georgetown.edu
Luciw-Dubas, Ulana, National Board of Medical Examiners, udubas@nbme.org
Ludovica, De Carolis, University of Milano-Bicocca, l.decarolis@campus.unimib.it
Luecht, Richard, rmluecht@uncg.edu
Lugu, Benjamin Kweku, University of Alabama, benjamin.lugu@aims.ac.rw
Lui, Angela, City University of New York, angela.lui@cuny.edu
Luo, Jinwen, UCLA, jevan.luo@gmail.com
Luo, Lin, Faculty of Education, The University of Hong Kong, Hong Kong, China, u3618120@connect.hku.hk
Lyons, Susan, Lyons Assessment Consulting, susan@lyonsassessment.com
Lyu, Weicong, University of Macau, weiconglyu@um.edu.mo
Ma, Jing, University of Iowa, majing@uiowa.edu
Ma, Wanjing (Anya), wanjingm@stanford.edu
Ma, Wenchao, University of Minnesota, wma@umn.edu
Ma, Ye, xxmlearn@gmail.com
Macias, Andrea, University of Southern California, andreajm@usc.edu
MacRae, Heather, heather.macrae@cambiumassessment.com
Madison, Matthew, University of Georgia, mjmadison@uga.edu
Maeda, Hotaka, Smarter Balanced, hotaka.maeda@gmail.com
Maeda, Yukiko, Purdue University, ymaeda@purdue.edu
Majerus, Mary, Missouri Department of Elementary and Secondary Education, mary.majerus@dese.mo.gov
Makinde, Henry, hsmakinde@uncg.edu
Malatesta, Jaime, GMAC, jmalatesta@gmac.com
Man, Kaiwen, University of Alabama, kman@ua.edu
Mann, Danielle, Duolingo, danimann@duolingocontractors.com
Mao, Tianzheng, Boston College, mao289765@gmail.com
Mao, Xia, xmao@nbome.org
Marder, Andrew, Caveon Test Security, andrew.marder@caveon.com
Mardones-Segovia, Constanza, University of Georgia, cam04214@uga.edu
Margolis, Melissa, NBME, mmargolis@nbme.org
María Leonor, Varas, DEMRE - Universidad de Chile, mlvaras@u.uchile.cl
Marini, Jessica, College Board, jmarini@collegeboard.org
Marion, Scott, Center for Assessment, smarion@nciea.org
Marquine, Maria, maria.marquine@duke.edu
Martin de los Santos Kleinz, Lisa, Johannes Gutenberg University Mainz, limartin@uni-mainz.de
Martineau, Joseph, ETS, jmartineau@ets.org

Martinez, Alfonso, Fordham University, alfonso.martinez@fordham.edu
Mason, Jessica, American Institutes for Research, jmason@air.org
Mastrup, Kayce, California State University, Monterey Bay, kmastrup@csumb.edu
Mattern, Krista, National Board of Medical Examiners, kburrus@nbme.org
Maul, Andrew, andrew.maul@ucsb.edu
Maur, Andreas, anmaur@uni-mainz.de
Maxfield, Maura, moriordan@umass.edu
Mazzullo, Elisabetta, University of Alberta, mazzullo@ualberta.ca
McBride, Yuanyuan, Pearson, malena.mcbride@pearson.com
McCaffrey, Daniel, ETS, dmccaffrey@ets.org
McClure, Kyla, University of Colorado Boulder, Kyla.Mcclure@colorado.edu
McFadden, Mara, James Madison University, mcfad2me@jmu.edu
McGee, Steven, The Learning Partnership, mcgee@lponline.net
McGrane, Joshua, University of Melbourne, joshua.mcgrane@unimelb.edu.au
McMillan, James, Virginia Commonwealth University, jhmcmill@vcu.edu
McNamara, Danielle, Arizona State University, dsmcnama@asu.edu
McVey, Jill, ACT, Inc., jill.mcvey@act.org
McWilliams, Kara, National Council of Bar Examiners, kara.n.mcwilliams@gmail.com
Mehou, Juste, James Madison University, mehoujm@dukes.jmu.edu
Mehrabi, Amirreza, Purdue University, amehrabi@purdue.edu
Mehrazmay, Roghayeh, mehr1365@live.com
Meijer, R.R., University of Groningen, r.r.meijer@rug.nl
Mena, Fernando, University of Massachusetts Amherst, fmenaserrano@umass.edu
Mendez, Jose, UC Irvine, jamend10@uci.edu
Mendoza, Guadalupe, UC Irvine, guadalam@uci.edu
Meng, Huijuan, Amazon Web Services, huijuam@amazon.com
Meng, Lionel, University of Wisconsin - Madison, lhmeng@wisc.edu
Merkle, Ed, University of Missouri, merklee@missouri.edu
Merkt, Martin, German Institute for Adult Education, merkt@die-bonn.de
Mese, Can, Indiana University, canmese@iu.edu
Meyer, Robert, Education Analytics & University of Wisconsin at Madison, rob.meyer.analytics@gmail.com
Michaelides, Michalis, University of Cyprus, michaelides.michalis@ucy.ac.cy
Michaels, Hillary, HumRRO, hillarymichaels@gmail.com
Michaluk, Lynette, Western Virginia University, lynnette.michalukk@mail.wvu.edu
Michel, Rochelle, rochelle.michel@gmail.com
Middlestead, Andrew, Michigan Department of Education, MiddlesteadA@michigan.gov
Middleton, Kyndra, Howard University, kvmiddleton@gmail.com
Miles, Julie, College Board, statsdocjam@gmail.com
Minchen, Nathan, nminchen@cainc.com
Ming, Norma, Carnegie Foundation for the Advancement of Teaching, nming@carnegiefoundation.org
Miratrix, Luke, Harvard Graduate School of Education, lmiratrix@g.harvard.edu
Mireles, Nicolas, James Madison University, mirelene@jmu.edu
Mislevy, Robert, University of Maryland, rmislevy@umd.edu
Missall, Kristen, University of Washington, kmissall@uw.edu
Mohammadi, Hamid, University of Alberta, hamid4@ualberta.ca
Mojoyinola, Mubarak, mubarak-mojoyinola@uiowa.edu
Molerov, Dimitri, Johannes Gutenberg University Mainz, dmolerov@uni-mainz.de
Moore, Steven, Carnegie Mellon University, StevenMo@andrew.cmu.edu
Morales, Amy, National Board of Medical Examiners, amorales@nbme.org
Morales, María Cristina, DEMRE - Universidad de Chile, maria.morales.za@uchile.cl
Morell, Linda, University of California, Berkeley, lindamorell@berkeley.edu
Morphew, Jason, Purdue University, jmorphew@purdue.edu
Morrison, Kristin, KMorrison@cainc.com
Moses, Tim, Buros Center for Testing, tmoses2@unl.edu

Moteane, Malitšitso, Winston-Salem State University, moteanem@wssu.edu
Mottesi, Gabriela, WestEd, gmottesi@wested.org
Moulder, Brad, bmoulder@ets.org
Mulcaire, Phoebe, Duolingo, phoebe@duolingo.com
Muller, Eric, American Institutes for Research, emuller@air.org
Munevver, Ilgun Dibek, TED University, munevver.ilgun@tedu.edu.tr
Muntean, William, National Council of State Boards of Nursing, wmuntean@ncsbn.org
Murphy, Sarah, Council For Professional Recognition, sarahm@cdacouncil.org
Murray, Amelia, University of Kansas, amelia.murray@ku.edu
Murray, Angela, University of Kansas, akmurray@ku.edu
Murray, Tonya, Stanford Graduate School of Education, tonyamur@stanford.edu
Muszyński, Marek, Institute of Philosophy and Sociology of the Polish Academy of Sciences, marek.muszynski@ifispan.edu.pl
Myoung, Eunjung, Stanford University, ejmyoung@stanford.edu
Nagasawa, Mirai, mnagasawa@crimson.ua.edu
Nagel, Marie, marie.nagel@uni-mainz.de
Nagy, Gabriel, IPN – Leibniz-Institute for Science and Mathematics Education, nagy@leibniz-ipn.de
Naim, Alper, ETS, nalper@ets.org
Nalbandyan, Roza, rn2407@stanford.edu
Nam, Sungjin, ACT, Inc, namsungjinn@gmail.com
Nash, Brooke, University of Kansas, bnash@ku.edu
Nason, Kelsey, James Madison University, nasonkt@jmu.edu
Nathan, Rebecca, Aviellah Curriculum Consulting, aviellahCC@gmail.com
Nesbitt, Jaylin, jnesbit2@wested.org
Nese, Joseph F. T., University of Oregon, jnese@uoregon.edu
Neumann, M., Free University of Amsterdam, m.neumann@vu.nl
Newton, Steve, Amplify Education Inc, snewton@amplify.com
Ngo, Trung, Northeastern University Seatle, trungnqcs2000@gmail.com
Nguyen, Hanh, hanh.nguyen@edmentum.com
Nguyen, Hiep, Eastern Illinois University, hmnguyen@eiu.edu
Nguyen, Tram-Anh Tran, University of Massachusetts, Amherst, anhn@umass.edu
Ni, Congning, Vanderbilt University, congning.ni@vanderbilt.edu
Nickodem, Kyle, University of Minnesota - Twin Cities, nicko013@umn.edu
Niessen, A.S.M, University of Groningen, a.s.m.niessen@rug.nl
Nissen, Jayson, LASSO, jayson.nissen@gmail.com
Nitkin, David, Transcend Education, david@transcendeducation.org
Nolan, Kate, Amira Learning, NolanKatherineA@gmail.com
North, Kai, George Mason University, knorth8@gmu.edu
Novak, Jennifer, Pearson, jennifer.novak@pearson.com
Nydick, Steven, Duolingo, swnydick@gmail.com
O'Donnell, Francis, fodonnell@nbme.org
O'Grady, Christina, cogrady@mainevirtualacademy.org
O'Reilly, Tenaha, toreilly@ets.org
O'Rourke, Kevin, University of Massachusetts Amherst, lofine@gmail.com
Oakley, Alexis, alexis-c-oakley@uiowa.edu
Ober, Teresa, ETS, tober@ets.org
Oberle, Catherine, Catherine.Oberle@edmentum.com
Oberle, Kristina, National Board of Medical Examiners, koberle@nbme.org
Obot, Patrick, University of Houston, patrickfridayobot@gmail.com
Odemwingie, Michelle, Achievement Network, modemwingie@achievementnetwork.org
Oh, Hyeonjoo, Riverside Insights, joannehj@gmail.com
Oh, Hyunjee, Teachers College, Columbia University, hyunjeeoh1223@gmail.com
Oliva, Michelle, EduDream, michelle@edudream.org
Olivera Aguilar, Margarita, American Institutes for Research, margarita.olag@gmail.com
Oliveri, Maria, Purdue University, oliveri.m@live.com

Olney, Andrew, University of Memphis, aolney@memphis.edu
Olson, Jacinta, jacinta.olson18@gmail.com
Olson-Rickel, Cara, ATLAS, University of Kansas, cara.rickel@ku.edu
Ong, Thai, NBME, tong@nbme.org
Oostrom, Janneke, Tilburg University, j.k.oostrom@tilburguniversity.edu
Ormerod, Christopher, christopher.ormerod@gmail.com
Orrill, Chandra, Rethink Learning Labs, chandra.orrill@rethinklearning.com
Osaro, Sarah, University of Iowa, sarahosaro24@gmail.com
Oyeniran, Daniel, The University of Alabama, dooyeniran@crimson.ua.edu
Ozkan, Fatih, fatih_ozkan1@baylor.edu
Palermo, Corey, Measurement Incorporated, cpalermo@measinc.com
Pan, Qianqian, National Institute of Education, Nanyang Technological University, qianqian.pan@nie.edu.sg
Pan, Tianshu, tianshu.pan@pearson.com
Pan, Yanjun, Southern Methodist University, yanjunp@mail.smu.edu
Panadero, Ernesto, Dublin City University, ernesto.research@gmail.com
Panzer, Jakob, Michigan State University, jakobspanzer@gmail.com
Papanastasiou, Elena, University of Nicosia, papanastasiou.e@unic.ac.cy
Pardos, Zachary, University of California, Berkeley, pardos@berkeley.edu
Park, Chan, University of Illinois Urbana-Champaign, parkchan@illinois.edu
Park, Hyemin, UC Berkeley, hyemin.park@berkeley.edu
Park, Junhee, University of Iowa, junhee-park@uiowa.edu
Park, Mi Jin, Graduate Center, City University of New York, mijinpark314@gmail.com
Park, Nohwon, University of Georgia, nohwon.park@uga.edu
Park, Seohee, American Board of Internal Medicine, hee6904@gmail.com
Park, Sojung, University at Albany, SUNY, spark42@albany.edu
Park, Yena, Duolingo, yena@duolingo.com
Park, Yookyeong, University at Albany, SUNY, ypark8@albany.edu
Parker, Emily, eparker@apertureed.com
Pastor, Dena, James Madison University, pastorda@jmu.edu
Patelis, Thanos, University of Kansas, patelis@ku.edu
Patterson, Chris, christopher-patterson@uiowa.edu
Patterson, Jaylene, University of Kentucky, jaylene.patterson@uky.edu
Patterson, Luke, American Institutes for Research, lpatterson@air.org
Patz, Richard, University of California, Berkeley, rpatz@berkeley.edu
Pei, Bo, University of South Florida, bpei@usf.edu
Pei, Francesca, University of California, San Francisco, francesca.pei@ucsf.edu
Pek, Jolynn, The Ohio State University, pek.5@osu.edu
Perie, Marianne, Measurement in Practice, mp@measurementinpractice.com
Perkoff, Elana Margaret, University of Colorado Boulder, emperkoff@gmail.com
Perna, Megan, Curriculum Associates, mperna@cainc.com
Pezeshki, Maryam, Cambium Assessment Inc., maryam.pezeshki@cambiumassessment.com
Pham, Duy, New Meridian, phamduy.edu@gmail.com
Phelps, Geoffrey, ETS, gphelps@ets.org
Phenow, Aurore, aurore.phenow@gmail.com
Phillips, Gary, Cambium Assessment, gary.phillips@cambiumassessment.com
Pilner Blair, Kristen, Stanford University, kpilner@stanford.edu
Pinto, Wallace, wallacenpj@gmail.com
Pitoniak, Mary, ETS, mpitoniak@ets.org
Plisiecki, Hubert, Institute of Philosophy and Sociology of the Polish Academy of Sciences, hplisiecki@gmail.com
Poe, Mya, Northeastern University, m.poe@northeastern.edu
Pokropek, Artur, IFiS PAN, artur.pokropek@gmail.com
Polikoff, Morgan, University of Southern California, polikoff@usc.edu
Poole, Frederick, Michigan State University, poolefre@msu.edu
Por, Han-Hui, ETS, HPor@ets.org

Porcelli, Victor, MDRC, victor.porcelli@mdrc.org
Potgieter, Cornelis, Texas Christian University, c.potgieter@tcu.edu
Powell, Sarah, The University of Texas at Austin, srpowell@utexas.edu
Powers, Sonya, sopowers@gmail.com
Prendergast, Caroline, cprendergast@absurgery.org
Prewett, Sara, prewettsara@gmail.com
Proctor, Thomas, tproctor@collegeboard.org
Purpura, David, Purdue University, purpura@purdue.edu
Pusey, Kerry, University of Pennsylvania, Kpusey@upenn.edu
Qiao, Xin, University of South Florida, xqiao@usf.edu
Qin, Qi, Cognia, qinqi715@gmail.com
Qiu, Chen, ACT, chenqiu_qpm.uky2022@outlook.com
Quanbeck, Mari, quanb016@umn.edu
Quesen, Sarah, WestEd, sarah.quesen@gmail.com
Quirk, Victoria, University of Illinois at Urbana-Champaign, vquirk3@illinois.edu
Rai, Snigdha, Kaplan North America, snigdha.rai@kaplan.com
Rajeb, Mehdi, mrajeb@crimson.ua.edu
Ramazan, Onur, Washington State University, onur.ramazan@wsu.edu
Rambis, Natalie, Menlo Education Research, nrambis@menloedu.org
Randall, Jennifer, University of Michigan, jennrand@umich.edu
Ranieri, Linda, NBME, lranieri@nbme.org
Ravand, Hamdollah, Vali-e-Asr University of Rafsanjan, ravandhamdollah@gmail.com
Rayfield, Norah, Education Quality and Accountability Office, norah.rayfield@eqao.com
Razavi, Pooya, Edmentum, pooya.razavi@protonmail.com
Read, Daniel, LOR Foundation, daniel@lorfoundation.org
Read, Wyatt, University of Delaware, wread@udel.edu
Reckase, Mark, Psychometric Solutions, reckase@msu.edu
Ren, He, University of Washington, heren@uw.edu
Rewley, Kelly, krewley@abim.org
Rezayi, Saed, NBME, srezayidemne@nbme.org
Richie-Halford, Adam, Graduate School of Education, Stanford University, adamrh@stanford.edu
Rijmen, Frank, Cambium Assessment, frank.rijmen@cambiumassessment.com
Rikoon, Samuel, American Institutes for Research, srikoon@air.org
Riley-Lepo, Erin, The College of New Jersey, rileylee@tcnj.edu
Rios, Joseph, Center for Measurement Justice, jrios@measurementjustice.org
Rios, Kristina, California State University, Fresno, krios@mail.fresnostate.edu
Rios, Oscar, University of California - Davis, osrios@ucdavis.edu
Ritter, Steven, Carneg, sritter@carnegielearning.com
Ritthaler, Mike, Pearson, mike.ritthaler@pearson.com
Roberts, Kari, Florida State University, klroberts@fsu.edu
Robertson, Maya, Grambling State University, mrobert5@gsumail.gram.edu
Robin, Frederic, ETS, frobin@ets.org
Robitaille, Jennifer, Riverside Insights, JRobitaille@ApertureEd.com
Rodriguez, Jennifer, University of Michigan, jaracely@umich.edu
Rodriguez, Michael, University of Minnesota, mcrdz@umn.edu
Rodríguez, Nelson Andrés, Instituto Colombiano para la Evaluación de la Educación - ICFES, nrodriguez@icfes.gov.co
Roeber, Edward, Michigan Assessment Consortium, ed.roeber22@gmail.com
Roeters-Solano, Heather, Pearson, heather.roeters-solano@pearson.com
Rolla, Andrea, arolla@wested.org
Rome, Logan, Curriculum Associates, lrome@cainc.com
Rosales de Veliz, Leslie Vanessa, JML Measurement and Testing Services, LLC, leslie.rosales@gmail.com
Rosca, Oxana, University at Albany, SUNY, orosca@albany.edu
Roussos, Louis, Cognia, roussos.louis1@gmail.com
Rozman, Mojca, IEA Hamburg, mojca.rozman@iea-hamburg.de

Ruiz-Primo, Maria, Stanford University, aruiz@stanford.edu
Ruiz Diaz, Shona, ETS, sruizdiaz@ets.org
Runge, Andrew, Duolingo, arunge@duolingo.com
Runyon, Christopher, National Board of Medical Examiners, CRunyon@nbme.org
Rupp, Andre, Center for Assessment, arupp@nciea.org
Russell, Javarro, Association of American Medical Colleges (AAMC), jrussell@aamc.org
Russell, Michael, Boston College, michael.russell@bc.edu
Rutkowski, Leslie, Indiana University, lrutkows@iu.edu
Rutzler, Erica, Association of American Medical Colleges, erutzler@aamc.org
Ryan, Ève, University of Alaska Fairbanks, eryan9@alaska.edu
Saavedra, Ana, Graduate School of Education, Stanford University, anamar@stanford.edu
Saavedra, Cristian, DEMRE - Universidad de Chile, crisaavedra@uchile.cl
Sabatini, Peter, Pearson, peter.sabatini@pearson.com
Sachdeva, Ruchi, Pearson, rujs77@gmail.com
Sahakyan, Narék, University of Wisconsin - Madison / WIDA, sahakyan@wisc.edu
Saiar, Amin, Board of Pharmacy Specialties, asaiar@aphanet.org
Sajidin, Zahra, Pearson, zahra.sajidin@pearson.com
Sakovich, Jeremy, Law School Admission Council, jsakovich@lsac.org
Saldivia, Luis, ETS, lsaldivia@ets.org
Salzer, Rebecca, MDRC, rebecca.salzer@mdrc.org
Samant, Yashad, Pearson, yashad.samant@gmail.com
Sanchez, Edgar, suppression305@gmail.com
Sanchez, Sergio, University of California, Davis, slsanchez@ucdavis.edu
Sandoval Hernandez, Andres, Department of Education, University of Bath, A.Sandoval@bath.ac.uk
Santelices, Maria, vsanteli@hotmail.com
Sarac, Merve, College Board, msarac@collegeboard.org
Sarig, Joni, National Institute for Testing & Evaluation (NITE), Joni@nite.org.il
Satkus, Paulius, psatkus@gmac.com
Sato, Edynn, Sato Education Consulting LLC, edynn@satoeducationconsulting.com
Sauder, Derek, American Board of Internal Medicine, dsauder@abim.org
Savage, Corey, University of Glasgow, corey.savage@glasgow.ac.uk
Savoie, Renee, renee.savoie@ct.gov
Sayin, Ayfer, Gazi University, ayfersayin@gazi.edu.tr
Schellman, Madeline, Pearson, madeline.schellman@pearson.com
Schenke, Katerina, ETS, katerinaschenke@gmail.com
Schlomer, Gabriel, University at Albany, SUNY, gschlomer@albany.edu
Schmidt, Katie, College Board, katie.a.schmidt@outlook.com
Schneider, Christina, cschne5934@aol.com
Schneider, Wei, Cambium Assessment, wei.schneider@cambiumassessment.com
Schröter, Hannes, FernUniversität Hagen, hannes.schroeter@fernuni-hagen.de
Schumacker, Randall, The University of Alabama, rschumacker@ua.edu
Schwartz, Bob, National Conference of Bar Examiners, schwartz.bob@gmail.com
Sciolla, Bruna, DEMRE - Universidad de Chile, bruna.sciolla@uchile.cl
Seissus, Emilia, Institut d'études Politiques de Paris, emilia.seissusercilla@sciencespo.fr
Serrentino, Dean, Curriculum Associates, dserrentino@cainc.com
Sgammato, Adrienne, asgammato@ets.org
Shaffer, David, University of Wisconsin Madison, dws@education.wisc.edu
Shamsaldeen, Fajer, Kuwait University, fajer.shamsaldeen@ku.edu.kw
Shanahan, Marie-Claire, University of Calgary, mcshanah@ucalgary.ca
Sharpnack, James, Duolingo, james.sharpnack@duolingo.com
Shaw, Emily, College Board, eshaw@collegeboard.org
Shear, Benjamin, University of Colorado Boulder, benjamin.shear@colorado.edu
Shen, Yawei, yawei.shen@pearson.com
Shenavai, Kevin, Johannes Gutenberg University Mainz, kshenava@uni-mainz.de

Shepard, Lorrie, lorrie.shepard@colorado.edu
Sheppard Miller, Gayle, Lower Kuskokwim School District, Bethel, Alaska, gayle_miller@lksd.org
Sherer, David, Carnegie Foundation, dsherer@carnegiefoundation.org
Shermis, Mark, Performance Assessment Analytics, LLC, mshermis@gmail.com
Shi, Qingzhou, qshi7@crimson.ua.edu
Shillo, Roi, National Institute for Testing & Evaluation (NITE), roishillo@gmail.com
Shin, David, Pearson, cshin0803@gmail.com
Shin, Emily, University of Toronto, emilyeunhai@gmail.com
Shin, Jihye, Guwol Girl's Middle School, jh_shin@snu.ac.kr
Shin, MinJeong, ACT, Inc., shinm@act.org
Shin, Seungwon, University of Iowa, seungwon-shin-1@uiowa.edu
Shor, Yael, National Institute for Testing & Evaluation (NITE), yaelshor@gmail.com
Shukla, Kathan, Indian Institute of Management Ahmedabad, kathans@iima.ac.in
Shyyan, Vitaliy, vshyyan@wisc.edu
Siebert, Julian, Stanford University, jms312@stanford.edu
Sikali, Emmanuel, NCES, Emmanuel.Sikali@ed.gov
Simmering, Vanessa, ATLAS, University of Kansas, simmering@ku.edu
Simpson, mary Ann, Carnegie Learning, msimpson@carnegielearning.com
Sinharay, Sandip, ETS, ssinharay@ets.org
Sireci, Stephen, University of Massachusetts Amherst, sireci@acad.umass.edu
Sirganci, Gozde, Southern Methodist University, gozdesirganci@gmail.com
Siu, Amanda, University of Florida, amanda.siu@ufl.edu
Skerrett, Allison, University of Texas at Austin, ASkerrett@utexas.edu
Skorupski, William, Data Recognition Corporation, william_skorupski@yahoo.com
Slater, Sharon, ETS, SSlater@ETS.ORG
Sloan, Caroline, ATLAS, University of Kansas, carolinesloan@ku.edu
Small, Electra, MDRC, electra.small@mdrc.org
Smith, Jessalyn, DRC, jsmith@datarecognitioncorp.com
Smith, Nate, University of Massachusetts, Amherst, nrsmith@umass.edu
Snow, Erica, Roblox, esnow@roblox.com
Soland, James, University of Virginia, jgs8e@virginia.edu
Solano-Flores, Guillermo, Stanford University, gsolanof@stanford.edu
Somay, Su, National Board of Medical Examiners, ssomay@nbme.org
Someshwar, Shonai, University of North Carolina Greenboro, s_somesh@uncg.edu
Somsong, Sarunya, Srinakharinwirot University & Southern Methodist University, ssomsong@smu.edu
Sondermann, Christina, German Institute for Adult Education, sondermann@die-bonn.de
Song, Dan, The University of Iowa, dan-song@uiowa.edu
Song, Victoria, Fordham University, vickysongthinks@gmail.com
Song, Woori, woori@snu.ac.kr
Song, Yi, ETS, ysong@ets.org
Soni, Parth, Indian Institute of Management Ahmedabad, phd19parths@iima.ac.in
Sonnleitner, Philipp, Luxembourg Centre for Educational Testing, University of Luxembourg, philipp.sonnleitner@uni.lu
Southerland, Joshua, Pearson, josh.southerland@pearson.com
Soyoye, Olushola, University of Delaware, olusholasoyoye@gmail.com
Sparks, Jesse, ETS, jsparks@ets.org
Specht-Boardman, Ryan, Comptency-Based Education Network, ryan@c-ben.org
St. George, Joseph, ACT, joseph.st.george@act.org
Stacho, Ladislav, Department of Mathematics, Simon Fraser University, lvstacho@gmail.com
Stahlhut, Laura, Institut für Bildungsevaluation Zürich AG, laura.stahlhut@ibe-edu.ch
Stark, Stephen, University of South Florida, sestark@usf.edu
Steedle, Jeffrey, jtsteedle@gmail.com
Stegers-Jager, Karen, Health Academy Radboudumc, karen.stegers-jager@radboudumc.nl
Stethen, Kelley, ACS Ventures, kelleyrwheeler@gmail.com
Stevens, Mark, SAS, mdstevens@uncg.edu

Stirn, Yuanyuan, yuanyuan-jiang-1@uiowa.edu
Stout, William, University of Illinois at Urbana-Champaign (Emeritus), w-stout1@illinois.edu
Strietholt, Rolf, IEA Hamburg, rolf.strietholt@iea-hamburg.de
Student, Sanford, University of Delaware, srstu@udel.edu
Su, Yu, Curriculum Associates, ysu@cainc.com
Su, Yu-Lan, suyulan@gmail.com
Subhiyah, Raja, rsubhiyah@nbme.org
Suh, Hongwook, Cambium Assessment, hongwooks@gmail.com
Suh, Yon Soo, NWEA within HMH, yon.soo.suh@nwea.org
Suk, Youmi, Teachers College, Columbia University, ysuk@tc.columbia.edu
Sun, Sundance Zhihong, The University of Melbourne, zhihongs@student.unimelb.edu.au
Sun, Yulan, DEMRE - Universidad de Chile, yulan.sun@uchile.cl
Sundar, Kripa, EdTech Recharge, kripasunda@gmail.com
Sung, Youjin, University of Maryland, yjsung@umd.edu
Sunnassee, Devdass, University of North Carolina at Greensboro, d_sunnas@uncg.edu
Suolang, Deji, University of Michigan - Ann Arbor, d.suolang@gmail.com
Supovitz, Jonathan, University of Pennsylvania, jons@upenn.edu
Surman, Stacy, The University of Alabama, ssurman@ua.edu
Svetina Valdivia, Dubravka, Indiana University, dsvetina@iu.edu
Swaby, Jeneve, wabyj@bc.edu
Sweeney, Sandra, Cognia, sandra.sweeney@cognia.org
Sweet, Tracy, University of Maryland, tsweet@umd.edu
Swinburne Romine, Russell, University of Kansas, rsr@ku.edu
Swinth, Emily, MDRC, emily.swinth@mdrc.org
Szendey, Olivia, oszende@wested.org
Tajik, Elham, Florida State University, et22e@fsu.edu
Talreja, Vinita, vgtalrej@amazon.com
Tan, Rachael, University of Colorado School of Medicine, rachael.tan@cuanschutz.edu
Tang, Cheng, The University of Georgia, cheng.tang@uga.edu
Tang, Nai-En, National Board of Chiropractic Examiners, naientang@gmail.com
Tang, XiaoDan, Northwestern University Feinberg School of Medicine Department of Medical Social Sciences, xiaodan.tang@northwestern.edu
Tang, Xiuxiu, University of Notre Dame, xtang8@nd.edu
Tang, Yuefeng, College Board, mayflowertang@gmail.com
Tao, Shuqin, Cambium Assessment Inc., shuqin.tao@cambiumassessment.com
Taub, Deborah, OTL Education Solutions, drdtaub@gmail.com
Tavares, Stephen, University of Virginia, st4yk@virginia.edu
Taylor, Catherine, ctaylor@uw.edu
Taylor, Christopher, Amazon Web Services, etaylorc@amazon.com
Taylor, Melinda, taylor.melinda@gmail.com
Templin, Jonathan, University of Iowa, jonathan-templin@uiowa.edu
Tenison, Caitlin, ETS, ctenison@ets.org
Terao, Takahiro, The National Center for University Entrance Examinations, terao@rd.dnc.ac.jp
Thacker, Arthur, Human Resources Research Organization, athacker@humrro.org
Thier, Michael, International Baccalaureate Organization, University of Oregon, michael.thier@ibo.org
Thomas, Jay, ACT Education Corp, jay.thomas@act.org
Thomas Murphy, Ciara, University of California, Davis, cthomasmurphy@ucdavis.edu
Thompson, Jake, wjakethompson@gmail.com
Thum, Yeow Meng, yeow.thum@gmail.com
Thurlow, Martha, University of Minnesota, thurl001@umn.edu
Tighe, Elizabeth, Georgia State University, etighe@gsu.edu
Tingir, Seyfullah, stingir@amplify.edu
Toker, Turker, Faculty of Educational Sciences, Department of Educational Measurement and Evaluation.Usak University, turker.toker@usak.edu.tr

Tolan, Shari, ATLAS, University of Kansas, shari.tolan@ku.edu
Tolentino, Lissette, University of Florida, ltolen@ufl.edu
Tong, Ye, National Board of Medical Examiners, yetong@nbme.org
Topczewski, Anna, Law School Admissions Council, topczewski.anna@gmail.com
Torres Irribarra, David, Pontificia Universidad Catolica de Chile, davidtorres@uc.cl
Toton, Sarah, Caveon, sarah.toton@caveon.com
Townley-Flores, Carrie, Graduate School of Education, Stanford University, ctflores@stanford.edu
Toyama, Yukie, yukie.toyama@berkeley.edu
Tran, Jasmine, University of California, Irvine, jazzytran@gmail.com
Tran, Peter, University of Massachusetts, Amherst, petertran@umass.edu
Trierweiler, Lukas, Johannes Gutenberg University Mainz, trierweiler@uni-mainz.de
Trout, Nick, Michigan State University, troutnic@msu.edu
True, Taylor, taylortrue15@gmail.com
Tsai, Chia-Lin, University of Northern Colorado, chialin.tsai@unco.edu
Tse, Winnie Wing-Yee, Vretta Inc., winnie.tse@vretta.com
Tu, Dongbo, Jiangxi Normal University, tudongbo@aliyun.com
Tucker, Eric, Equity by Design, eric@equitybydesign.org
Turner, Ronna, University of Arkansas, rcturner@uark.edu
Twing, Jon, HumRRO, jon.s.twing@icloud.com
Ulitzsch, Esther, University of Oslo, ulitzsch@ipn.uni-kiel.de
Umeh, Emmanuel, University of Missouri, ecung6@missouri.edu
Ungashe, Albu, Stanford University, aungashe@stanford.edu
Vafis, Greg, ETS, gvafis@ets.org
Valdivia Medinaceli, Montserrat, Curriculum Associates, mvaldivia@cainc.com
van Buel, Sanjay, Erasmus University Rotterdam, vanbuel@essb.eur.nl
van Campenhout, Rachel, VitalSource, Rachel.VanCampenhout@vitalsource.com
Van Dusen, Ben, Iowa State University, bvd@iastate.edu
Van Orman, Dustin, Western Washington University, vanormd2@wwu.edu
van Rijn, Peter, ETS Global, pvanrijn@etsglobal.org
Vang, Gao, gao.vang@spps.org
Velez, Sophia, EduDream, sophia@edudream.org
Ventura, Claudia, claudia.j.ventura@uconn.edu
Vernon, Annette, University of Iowa, annette-vernon@uiowa.edu
Villafuerte, Catherina, University of Connecticut, catherina.villafuerte@uconn.edu
Villanueva, Kristin, UNC Charlotte, kvillan1@charlotte.edu
Villegas, Leslie, villegasl@newamerica.org
Vispoel, Walter, The University of Iowa, walter-vispoel@uiowa.edu
Vivas Corrales, Pablo, University of Minnesota, vivas014@umn.edu
Vo, Thao, Center for Measurement Justice, tvo@measurementjustice.org
Vo, Yen, yen-vo@uiowa.edu
von Davier, Alina, Duolingo, avondavier@duolingo.com
von Davier, Matthias, Boston College, vondavim@bc.edu
Vorenkamp, Ellen, Michigan Assessment Consortium, vorenke83@gmail.com
Vu, Lien, University of Delaware, lienvu@udel.edu
Vue, Kory, University of Minnesota, vuexx199@umn.edu
Wackerle-Hollman, Alisha, University of Minnesota, wacke020@umn.edu
Walker, Marcus, National Commission on Certification of Physician Assistants, marcusw@nccpa.net
Walker, Michael, HumRRO, memwalker@gmail.com
Walkington, Candace, Southern Methodist University, cwalkington@mail.smu.edu
Walkowiak, Temple, North Carolina State University, tawalkow@ncsu.edu
Wallace, Leslie, Westat, LeslieWallace@westat.com
Wan, Lei, College Board, lwan@collegeboard.org
Wan, Siyu, ABIM, siyuwan93@outlook.com
Wang, Aijun, wajlm2003@gmail.com

Wang, Bowen, University of Florida, bowen.wang@ufl.edu
Wang, Caroline, Education Analytics, cwang@edanalytics.org
Wang, Chun, University of Washington, wang4066@uw.edu
Wang, Chunxin, ACT, Inc., annwang728@gmail.com
Wang, Daxun, Jiangxi Normal University, wangda.xun@163.com
Wang, Dongwei, American Board of Surgery, dwang@absurgery.org
Wang, Fang, University of South Carolina, fw3@email.sc.edu
Wang, Haiyan, Pearson, haiyan.wang@pearson.com
Wang, Hongling, ACT, hongling.wang@act.org
Wang, Huijuan, huijuan@email.sc.edu
Wang, Kun, Rethink Learning Labs, kun.wang@rethinklearning.com
Wang, Kuo, Southern Methodist University, wangp@mail.smu.edu
Wang, Min, College Board, wangminemail@gmail.com
Wang, Qin, 364596226@qq.com
Wang, Shaojie, Guangdong University of Education, wang021112@126.com
Wang, Shichao, Shichao.Wang@act.org
Wang, Shiyu, University of Georgia, swang44@uga.edu
Wang, Tao, University of Victoria, taow@uvic.ca
Wang, Ting, szpku.grady@gmail.com
Wang, Xiyu, wang5608@purdue.edu
Wang, Xu, University of Michigan, xwanghci@umich.edu
Wang, Yu, New York University, yw3060@nyu.edu
Wang, Yuan, ETS, ywang@ets.org
Wang, Ze, awsz@amazon.com
Wang, Zhen, Hkeaa, zwang@hkeaa.edu.hk
Wang, Zhuoran, NCSBN, wzhranran@gmail.com
Wang, Zifan, Syracuse University, zwang345@syr.edu
Wang, Zuowei, ETS, zwang@ets.org
Warner, Jeffrey, Georgetown University School of Continuing Studies, jww77@georgetown.edu
Warner, Zachary, New York State Education Department, zachary.warner@nysed.gov
Watts, Field, ETS, fwattsjr@ets.org
Webb, Blue, American Institutes for Research, bwebb@air.org
Webb, Noreen, webb@ucla.edu
Weeks, Jonathan, American Institutes for Research, jweeks@air.org
Wei, Elizabeth, Cambium Assessment Inc., elizabeth.wei@cambiumassessment.com
Wei, Hsin-Ro, Riverside Insights, esep23@gmail.com
Wei, Junhuan, Jiangxi Normal University, weijunhuan5@163.com
Wei, Xiaobo, University of South Carolina, xiaobow@email.sc.edu
Wei, Xin, Digital Promise, xwei@digitalpromise.org
Weinstein, Marc, Caveon, marc.weinstein@caveon.com
Weiss, David, University of Minnesota, djweiss@umn.edu
Welch, Catherine, catherine-welch@uiowa.edu
Wells, Craig, University of Massachusetts Amherst, cswells@umass.edu
Welsh, Megan, welsh.megan@gmail.com
Wen, Hongbo, Beijing Normal university, whb@bnu.edu.cn
Wentzlof, Kelly, Stanford University, kelwentz@stanford.edu
Westrick, Paul, College Board, westrick_3@hotmail.com
Wheeler, Jordan, University of Nebraska - Lincoln, jwheeler21@unl.edu
Whitcomb, David, ATLAS, University of Kansas, d_whitcomb@ku.edu
White, Lauren, law03k@gmail.com
Whitmer, John, Federation of American Scientists, John@ld-insights.com
Wilberforce, Winifred, Ohio State University, wilberforce.2@osu.edu
Wild, Autumn, wildan@jmu.edu
Wiley, Andrew, ACS Ventures, LLC, awiley999@gmail.com

Will, Kelsey, University of Minnesota, will3559@umn.edu
Williams, Angie, amlong81@gmail.com
Williams, Kevin, ETS, kmwilliams@ets.org
Williamson, Benjamin, CYBER.ORG, benjamin.williamson@cyber.org
Wilson, Joshua, University of Delaware, joshwils@udel.edu
Wilson, Mark, University of California, Berkeley, MarkW@berkeley.edu
Wilson, Quinell, Cambium Assessment, quinell.wilson@cambiumassessment.com
Wind, Stefanie, The University of Alabama, swind@ua.edu
Wine, Marjorie, Accessible Teaching, Learning and Assessment Systems (ATLAS), University of Kansas, mwine122@gmail.com
Winter, Sonja, University of Missouri, sdwinter@missouri.edu
Wise, Steven, EngagedMeasurement, stevewise23@gmail.com
Witmer, Sara, Michigan State University, switmer@msu.edu
Wolfe, Edward, Pearson, ed.wolfe@pearson.com
Wollack, James, University of Wisconsin - Madison, jwollack@wisc.edu
Wong, George, Pearson, george.wong@pearson.com
Woo, Ada, Ascend Learning, ada.woo@atitesting.com
Wood, Scott, ACT, Inc., woods@act.org
Woolf, Sherri, Consultant, sherriwoolf@gmail.com
Worrell, Frank, frankc@berkeley.edu
Worthington, Michelle, ETS, mworthington@ets.org
Woulfin, Sarah, The University of Texas at Austin, sarah.woulfin@austin.utexas.edu
Wright, Daniel, dbrookswr@gmail.com
Wright, Keith, The Enrollment Management Association, kwright@enrollment.org
Wry, Erin, Boston College, erin.wry@bc.edu
Wu, Meng, Educational Testing Service, mwu@ets.org
Wu, Shushan, University of Georgia, Shushan.wu@uga.edu
Wu, Sirui, University of British Columbia, sirui.wu@ubc.ca
Wu, Tong, tong.wu@riversideinsights.com
Wu, Xiaopeng, Faculty of Education, The University of Hong Kong, Hong Kong, China, wuxp722@hku.hk
Wu, Yi-Chen, National Center on Educational Outcomes, University of Minnesota, wuxx0207@umn.edu
Wu, Yi-Fang, Cambium Assessment, Inc., Yi-Fang.Wu@cambiumassessment.com
Wu, Zebing, Cambium Assessment, zebing.wu@cambiumassessment.com
Wyatt, Jeff, College Board, jwyatt@collegeboard.org
Wylen, Mel, WestEd, mwylen@wested.org
Wylie, E. Caroline, Center for Assessment, ecwylie@nciea.org
Wyman, Austin, University of Notre Dame, awyman@nd.edu
Wyse, Adam, Renaissance, adam.wyse@renaissance.com
Xi, Nuo, ETS, nxi@ets.org
Xia, Yan, University of Illinois at Urbana-Champaign, yx18@illinois.edu
Xiao, Daihui, xiaodaih@msu.edu
Xiao, Min, University of Kentucky, min.xiao@uky.edu
Xiao, Xingyao, University of California, Berkeley, xiaoxg@berkeley.edu
Xiong, Jiawei, Curriculum Associates, jxiong@cainc.com
Xiong, Yao, Google, xyxiongyao@gmail.com
Xu, Gongjun, University of Michigan, gongjun@umich.edu
Xu, Guoliang, Teachers College, Columbia University, gx2150@tc.columbia.edu
Xu, Jiajun, Cambium Assessment, jiajun.xu@cambiumassessment.com
Xu, Jing-Ru, jingruxu2013@gmail.com
Xu, Rujun, University of Virginia, bwp7ab@virginia.edu
Xu, Xiaochen, University of California, Davis, xixu@ucdavis.edu
Xu, Yangmeng, University of Alabama, yxu81@crimson.ua.edu
Xu, Yingshi, Beijing Normal university, xuyingshicf@163.com
Xue, Mingfeng, University of California, Berkeley, mingfengxue@berkeley.edu

Yan, Duanli, ETS, DYan@ETS.ORG
Yan, Pengfei, University of Maryland, College Park, pyan12@umd.edu
Yancey, Kevin, Duolingo, kyancey@duolingo.com
Yaneva, Victoria, National Board of Medical Examiners, vyaneva@nbme.org
Yang, Ji Seung, University of Maryland, jsyang@umd.edu
Yaxuan Yang, University of Georgia, yyaxuan@uga.edu
Yang, Yi, ETS, yyang@ets.org
Yao, Lihua, Northwestern University, lihua.yao@northwestern.edu
Yarmohammadtoosky, Sahar, Kennesaw State University, yarmohamadishr@gmail.com
Yavuz, Sinan, American Institutes for Research, yavuzsinan@gmail.com
Ye, Sangbeak, Florida Atlantic University, sye@fau.edu
Yeatman, Jason, Stanford University, jyeatman@stanford.edu
Yeboah, Clement, The University of Southern Mississippi, clement.yeboah@usm.edu
Yee, Darrick, American Institutes for Research, dyee@air.org
Yen, Shu Jing, Center for Applied Linguistics, syen@cal.org
Yin, Yue, University of Illinois at Chicago, yueyin@uic.edu
Yoo, Hanwook, Ascend Learning, hanuki82@gmail.com
Young, Mackenzie, mackenzie.young@cambiumassessment.com
Young, Rebecca, Cognia, rebecca.young@cognia.org
Yu, Haoyang, University of Kansas, haoyang.yu@ku.edu
Yu, Kefan, ky285@uw.edu
Yuan, Ye, GMAC, yyuan@gmac.com
Yung, Sammy, University of Calgary, sammy.yung@ucalgary.ca
YuZhi, Yan, Jiangxi Normal University, yanyz0818@163.com
Zahner, Doris, Council for Aid to Education, dzahner@cae.org
Zapata-Rivera, Diego, ETS, dzapata@ets.org
Zapparrata, Nicole, Ascend Learning, nicole.zapparrata@ascendlearning.com
Zegers, Mónica, UCSF, Monica.zegerslarrain@ucsf.edu
Zeiser, Kristina, AIR, kzeiser@air.org
Zelikman, Eric, Stanford University, ezelikman@cs.stanford.edu
Zeng, Biao, Beijing Normal University, biaozeng@mail.bnu.edu.cn
Zeng, Ji, zengj@michigan.gov
Zenisky, April, University of Massachusetts, Amherst, azenisky@educ.umass.edu
Zhai, Xiaoming, University of Georgia, xiaoming.zhai@uga.edu
Zhang, Bo, University of Illinois at Chicago, bozhang3@illinois.edu
Zhang, Chuqing, Mt. Holyoke, zhang49c@mtholyoke.edu
Zhang, Fan, University of Delaware, fzhang@udel.edu
Zhang, Jihong, University of Arkansas at Fayetteville, z.john92@gmail.com
Zhang, Jinming, University of Illinois at Urbana-Champaign, jmzhang@illinois.edu
Zhang, Lijin, Stanford Graduate School of Education, lijinzhang@stanford.edu
Zhang, Mingqin, mqziowa@gmail.com
Zhang, Nan, University of Maryland, nanzhang@umd.edu
Zhang, Oliver, College Board, ozhang@collegeboard.org
Zhang, Susu, University of Illinois at Urbana-Champaign, szhan105@illinois.edu
Zhang, Tiejun, zhangtiejun310@gmail.com
Zhang, Ting, American Institutes for Research, tzhang@air.org
Zhang, Wei, Augusta University, WZHANG2@augusta.edu
Zhang, Xiaohang, joyz1990@byu.edu
Zhang, Xiuyuan, The College Board, xzhang@collegeboard.org
Zhang, Yichi, American Institutes for Research, yzhang97@usc.edu
Zhang, Yifan, The University of Hong Kong, zyf2020@connect.hku.hk
Zhang, Yu, The Federation of State Boards of Physical Therapy, yzhang@fsbpt.org
Zhang, Yuxiao, Purdue University, zhan3971@purdue.edu
Zhang, Zhiyong, University of Notre Dame, zzhang4@nd.edu

Zhao, Xinchu, Roblox, xinchuz@gmail.com
Zheng, Guoguo, Amplify Education Inc, gzheng@amplify.com
Zheng, Xiaying, American Institutes for Research, xzheng@air.org
Zheng, Yi, Arizona State University, yi.isabel.zheng@asu.edu
Zhong, Xiaoting, University of Iowa, xzhong6@uiowa.edu
Zhou, Tianyi, University of Maryland, tianyi@umd.edu
Zhou, Xinchang, xz77@illinois.edu
Zhou, Xuechun, greenwhite0619@gmail.com
Zhou, Yan, Houghton Mifflin Harcourt Publishing, yanzhouyzh@gmail.com
Zhou, Yiyun, yyzhou@nbme.org
Zhu, Sizheng, Roblox, sizhengzhu@outlook.com
Ziedzor, Reginald, Amplify Education Inc, ziedzor@hotmail.com
Zlatkin-Troitschanskaia, Olga, Johannes Gutenberg University Mainz, troitschanskaia@uni-mainz.de
Zopluoglu, Cengiz, University of Oregon, cengiz@uoregon.edu
Zou, Tongtong, University of Wisconsin - Madison, tz2345@tc.columbia.edu
Zu, Jiyun, ETS, jzu@ets.org
Zubenko, Vladimir, ETS, vzubenko@ets.org
Zwick, Rebecca, University of California, Santa Barbara, rzwick@cox.net

National Council on Measurement in Education is very grateful to the following organizations for their generous financial support of our 2025 Annual Meeting

PLATINUM

CollegeBoard

Pearson

GOLD

duolingo english test

Center for Assessment

SILVER

ACS Ventures — Bridging Theory & Practice

ACT®

cognia

AAMC — Medical College Admission Test®

DRC — Data Recognition Corporation

NBME®

edCount LLC — because all students count

Curriculum Associates — i-Ready

nwea — believe in what's possible

Cambium Assessment

Harvard Graduate School of Education

American Board of Internal Medicine®

FRIEND

BEAR CENTER — Berkeley Evaluation and Assessment Research Center

UCLA CRESST

UMass Amherst — College of Education

LANDYARDS

VECTOR PSYCHOMETRIC GROUP

ncme

National Council on Measurement in Education
19 Mantua Road | Mount Royal, NJ 08061 | 856.284.3700 | ncme@ncme.org
www.ncme.org