Moving Measurement Forward

LOS ANGELES

Welcome from the Program Chairs

Welcome to the 2026 NCME Annual Meeting!
We are thrilled to have you with us in Los Angeles this year.

The theme of the 2026 NCME Annual Meeting is Moving Measurement Forward. The field of educational measurement is rapidly evolving in response to emerging technologies, diverse learner needs, and shifting policy landscapes. In turn, it is has never been more important for researchers, practitioners, and policymakers to explore innovative approaches that advance the science and practice of measurement. As we look toward the future, it is clear that progress in our field does not happen in isolation. We grow stronger by learning from one another—through shared research, diverse perspectives, and real-world examples from practice. The collective dedication of the NCME community is what fuels innovation, inspires new ideas, and helps us all move measurement forward together.

Our intention with the 2026 Annual Meeting is to offer a program that inspires you to take meaningful actions that move measurement forward. We invite you to engage with this year's training sessions and in-person presentations through that lens. Our hope is that you leave the conference with fresh ideas to explore or new perspectives on how to advance your ongoing research and practice. The program is rich with sessions spanning many pillars of educational measurement, and we hope you find this year's presentations as energizing and thought-provoking as we do.

There are a few events of within the Annual Meeting that we wanted to highlight for you, as you review the program and decide which sessions to attend.

- NCME 2026 includes over 20 training sessions on **Wednesday, April 8th**. There will be half-day (4-hour) and mini (2-hour) sessions.
- All three conference presentation days from **April 9th-11th** offer full days of great sessions organized in numerous formats: Coordinated Paper Sessions, Organized Discussions, and Individual Paper Sessions. Additionally, there will be three types of eBoard sessions: Innovation Demonstration, Individual eBoards, and Graduate Student In-Progress Research eBoards. Finally, we are featuring a new session type this year, the Coordinated Poster Session, which is a hybrid of a traditional presentation format with an eBoard format.
- The first day, **April 9th**, 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 10th**, we can look forward to starting our day at the NCME Business Meeting and Presidential Address, beginning at 7:30 AM.
- The final presentation day, **Saturday April 11th**, 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 highlighted sessions that attendees might be interested in each day.

Thursday April 9th:

- Remembering Jim Popham, **7:45am-9:15am**
- In Memory of Neil Dorans: A Life Dedicated to Equity and Fairness in Assessment, **9:45am-11:15am**
- NCMEntoring: Fireside Chat (Membership Committee Session), **11:30am-1pm**
- Innovation Demonstrations session, **1:45-2:45pm**
- Assessment Needs: Can We Build Balanced Systems of Assessment to Address Them? (Classroom Assessment Committee session), **1:45pm-3:15pm**
- AI Ethics, Policy, and Practice in Training Measurement Professionals: A Panel Discussion (Educators of Measurement SIGIMIE session)

Friday April 10th:

- The Revision of Testing Standards: Foundations (Joint AERA-NCME Session), **9:45am-11:15am**
- NCME Career Achievement Award, Sandip Sinharay, Reflection on Reporting High-Quality Scores Using Statistics, Measurement, and Professional Judgment, **1:45-3:15pm**

WELCOME FROM THE PROGRAM CHAIRS

- Threats and Opportunities on The Federal Statistics Landscape, **1:45-3pm**
- Assessment to Promote Civic Learning: New Volume in the NCME Book Series (Invited Session), **3:30-5pm**
- Battle of the Presidents (Past and Future), **3:30-5pm**
- NCME Task Force Conversation **3:30-5pm**

Saturday April 11th:

- The Revision of Testing Standards: Operations (Joint AERA-NCME Session), **7:45-9:15am**
- The Revision of Testing Standards: Applications (Joint AERA-NCME Session), **9:45-11:15am**
- Beyond “National” Measurement: An Introduction to International Assessment Research and Practices, **9:45-11:15am**
- The Content and Psychometric Partnership: A Panel Discussion, **9:45-11:15am**
- In Memory of Robert J. Mislevy: A Legacy of Innovations in Measurement, **9:45-11:15am**
- An NCME Debate Session: The AI Landscape for Industry and Academia, **11:30am-12:45pm**
- Navigating NCME Journals: Pathways to Publication (Publications Committee Session), **11:30am-1:00pm**
- Psychometric Overlords Take 2: Prove Me Wrong (An Invited Coordinated eBoard), **1:45-3:15pm**

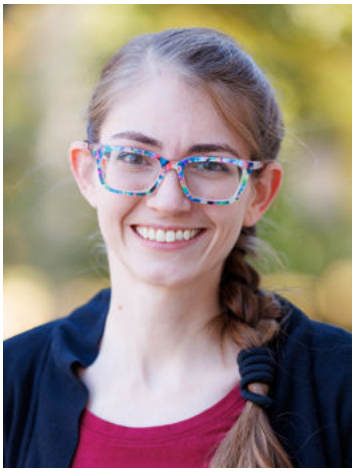
IIt takes a village to create such a stellar program, and we are so thankful to all who have contributed. This includes proposal authors, and those who volunteered to serve as Chair or Discussant. Finally, to the many who volunteered to assist with the review process—we are so grateful for your time and expertise. Thank you for helping to maintain the high quality of rigor of NCME. We recognize the names of all reviewers and invited review panels in this program.

We want to thank Joe Grochowalski and Wenchao Ma, the Training and Professional Development Committee Co-Chairs, as well as Kayla Burt and Claudia Ventura, the Graduate Student Issues Committee Co-Chairs, for all of their work on the program. We are grateful to the previous NCME program chairs Katherine Castellano and Scott Monroe, for their encouragement and support throughout the year. We appreciated the help, feedback and discussion with next year’s program chairs (Okan Bulut, Ikkyu Choi) and president (Kadriye Ercikan). We also want to thank both the former NCME Executive Director, Rich Patz, and current NCME Executive Director, Susan Lyons, for their support and feedback throughout the year. Finally, we want to thank the NCME President Amy Hendrickson for her continuous help, support, and encouragement, and for giving us this opportunity; it is an honor to serve NCME in this capacity.

We hope you enjoy the conference!

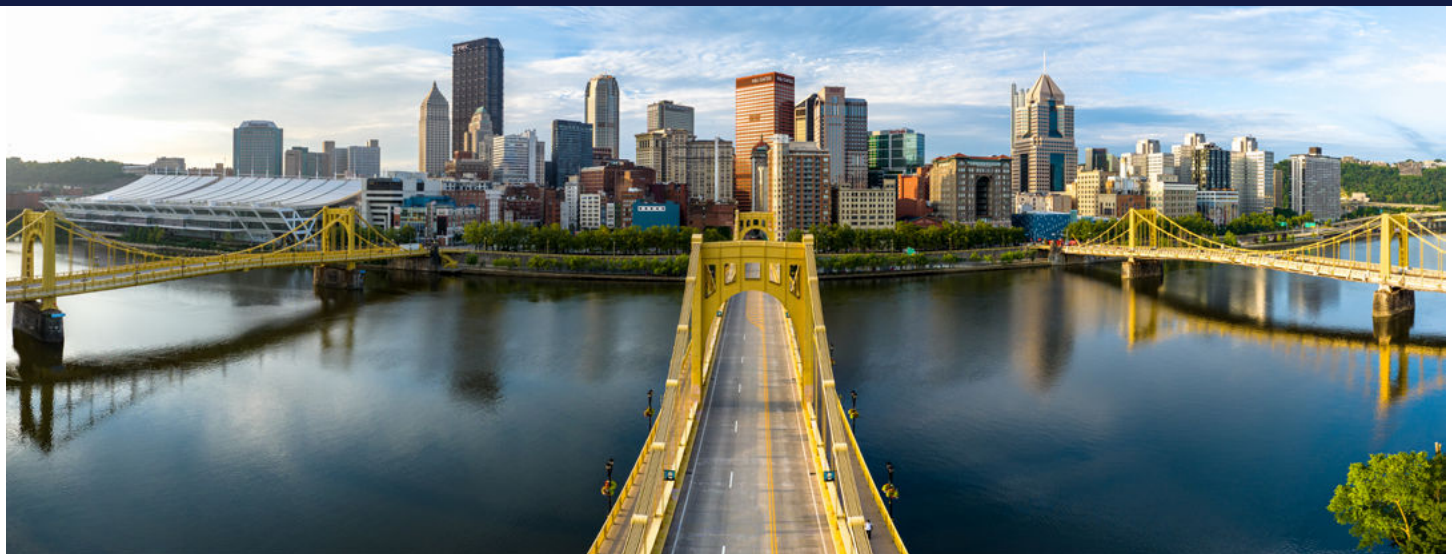
Pamela Kaliski and Stefanie A. Wind
2026 NCME Annual Conference Co-Chairs

Pamela Kaliski
2026 NCME Annual Conference Co-Chair



Stefanie A. Wind
2026 NCME Annual Conference Co-Chair

SAVE THE DATE!



AIME-Con 2026

October 5–7 | Pittsburgh

This year's conference theme, "Measurement Science in AI-Integrated Assessment and Pedagogy," explores the vital intersection of AI innovation and psychometric rigor, ensuring that the next generation of educational tools is anchored in validity, reliability, and fairness.

AIME-Con brings together an interdisciplinary group of leading experts in AI, psychometrics, education, NLP, and learning analytics to bridge gaps between these fields.

Call for sessions will open April 1.

The submission deadline is June 14.

AIME-CON 2026

Pittsburgh, PA, USA ncme

https://ncme.org/events/aime-conference/

SAVE THE DATE!

2026 Classroom Assessment Conference

November 12–14 | Bloomington, Minnesota

NCME is excited to announce that the 6th NCME Conference on Classroom Assessment will take place November 12–14, 2026, at the Radisson Blu, located in the Mall of America in Bloomington, Minnesota. This year's conference promises to be a dynamic and engaging event focused on advancing effective classroom assessment practices across educational contexts.

This gathering will unite a broad spectrum of voices—including researchers, K–12 educators, youth leaders, and Indigenous communities—to explore innovative approaches that center equity, relevance, and learner empowerment in assessment practices. Attendees can expect opportunities to collaborate, share best practices, and shape the future of classroom assessment through research, practical tools, and policy-relevant conversations.

https://ncme.org/events/classroom-assessment-conference/

TABLE OF CONTENTS

SCHEDULE AT-A-GLANCE

FULL SCHEDULE

For the most up to date schedule, use the app.

Any additions/changes to the schedule will also be posted in the document linked here.

2026 PROGRAM UPDATES

GENERAL MEETING INFORMATION

Welcome to the 2026 NCME Annual Meeting in Los Angeles!

NCME/AERA REGISTRATION & INFORMATION DESK

If you have already registered for the NCME and/or AERA conference, you may check-in and print out your conference badge at the NCME Information Desk, located on Podium 5th Floor in the Wilshire Grand pre-function area at the Inter-Continental Los Angeles Downtown. If you still need to register, you will need to visit the AERA booth in the exhibit hall, located in the L.A. Convention Center.

On-site registration will only be available at the Convention Center, which is a mile away from the NCME hotel, so we strongly recommend registering before you travel. Badge pick up is available at the NCME hotel, but only for those are registered in advance.

NCME registration will be open the following hours:

- Wednesday, April 8, 7:30 a.m.–5 p.m.
- Thursday, April 9, 7:30 a.m.–5 p.m.
- Friday, April 10, 7:30 a.m.–5 p.m.
- Saturday, April 11, 7:30 a.m.–4 p.m.

NCME GIVES BACK

Donate to the Los Angeles Education Partnership (LAEP) which has been transforming education since 1984. LAEP's story is one of determination and action. In 1983, inspired by the call to action in "A Nation at Risk," David Abel challenged Peggy Funkhouser to make a difference. The result was LAEP. Since our inception, we've been dedicated to working with the community. And we are proud to share that LAEP is the first nonprofit in Los Angeles to focus on educational equity.

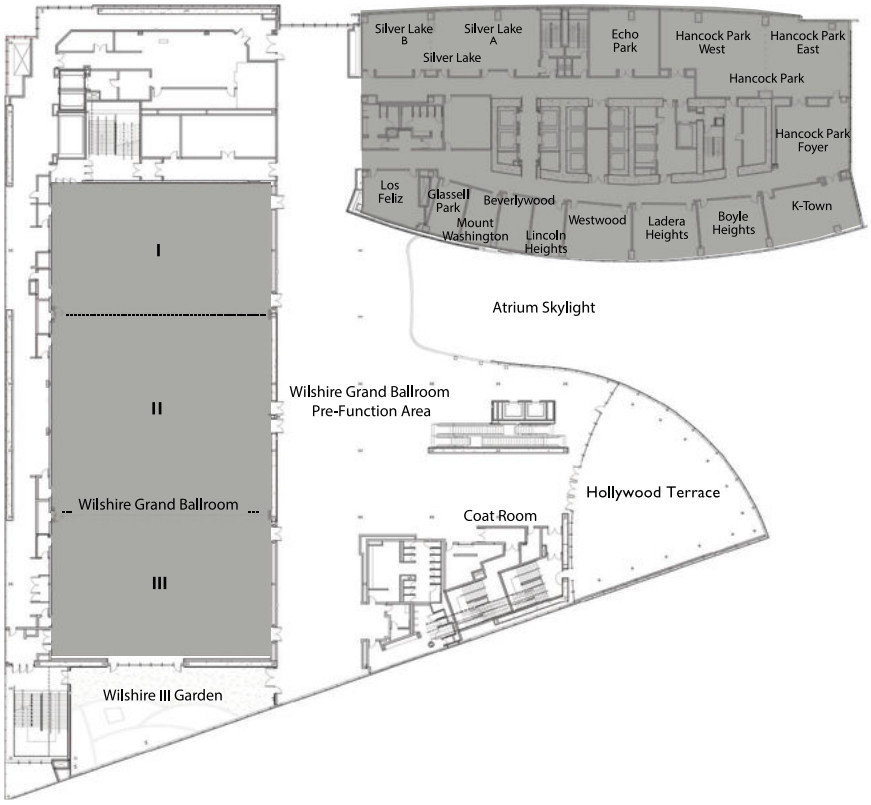
LOS ANGELES EDUCATION PARTNERSHIP

DOWNLOAD THE OFFICIAL NCME EVENTS APP

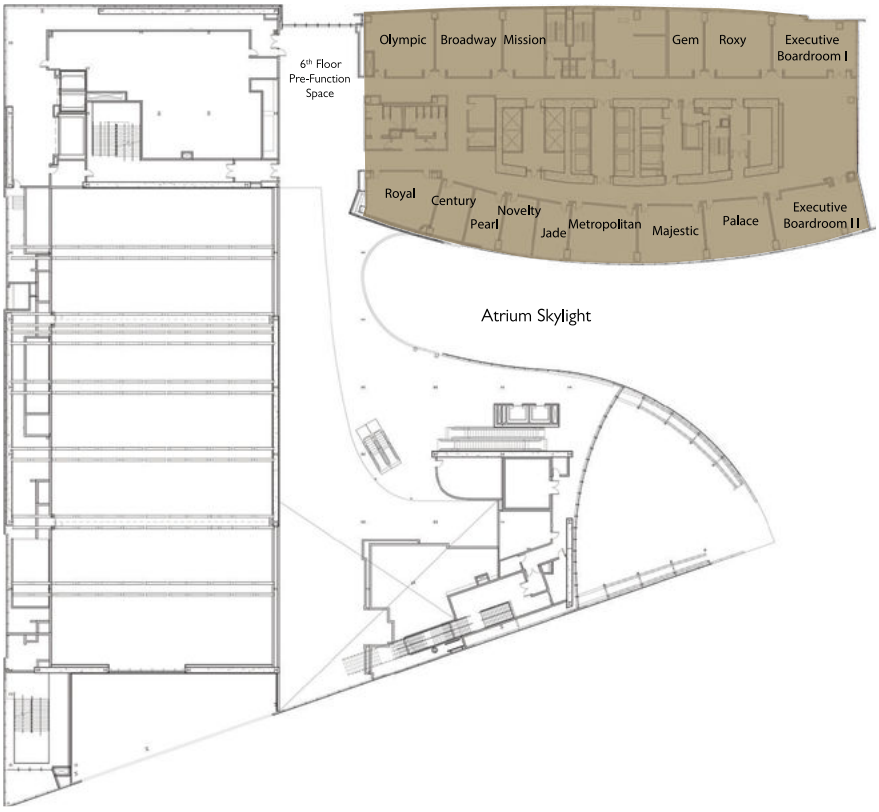
- Download the eConference.io app from the App Store or Google Play.
- Open the app and enter NCME26 as the conference code.
- Once you are in the app, "Click to Log In" at the top of the screen and enter the log-in credentials from the Know Before You Go email to access (i.e. the email with which you registered).

FLOOR PLANS

PODIUM 5 TH FLOOR



PODIUM 6 TH FLOOR



FLOOR PLANS

PODIUM 7 TH FLOOR

STAY CONNECTED

#NCME26

LinkedIn: ncme38

BlueSky: @ncme38

Facebook: NCMEPage

X/Twitter: @ncme38

NCME LEADERSHIP

NCME OFFICERS

President — **Amy Hendrickson**, *College Board*

President-Elect — **Kadriye Ercikan**, *Educational Testing Services*

Past President — **Andrew Ho**, *Harvard University*

NCME DIRECTORS

Leslie Keng
National Board of Medical Examiners

Brian Leventhal
James Madison University

Susan Lottridge
Pearson

Rochelle Michel
Smarter Balanced

Brad McMillen
Wake County Public School System

Zach Warner
New York State Education Department

NCME EXECUTIVE DIRECTOR

Susan Lyons

EDITORS

Journal of Educational Measurement
Won-Chan Lee
University of Iowa

Chinese/English Journal of Educational Measurement and Evaluation
Tao Xin
Beijing Normal University

Yi Zheng
Arizona State University

Educational Assessment
Alison Bailey & Felipe Martinez
UCLA

Educational Measurement: Issues and Practice
Heather Buzik
ACT

ITEMS
Stella Kim
University of North Carolina–Charlotte

NCME Book Series Chair
Kadriye Erkican
Educational Testing Services

News@NCME Editor
Cheng Hua
University of Montevallo

2026 ANNUAL MEETING CHAIRS

Annual Meeting Program Chairs

Pamela Kaliski
American Board of Internal Medicine

Stefanie Wind
University of Alabama

Training and Professional Development Committee Chairs

Joe Grochowalski
College Board

Wenchao Ma
University of Minnesota

Graduate Student Sessions Chairs

Kayla Burt
SUNY–Buffalo

Claudia Ventura
University of Connecticut

Fitness Run/Walk Director

Katherine Castellano
Educational Testing Services

GRADUATE STUDENT POSTER COMPETITION JUDGES

Valerie Ofori
Susan Lyons
Burcu Arslan
Magdalen Beiting-Parrish
Brian Leventhal
Monste Medinaceli
Jon Beard
Justin Kern
Tukhbita Nawmi
Jiawei Xiong
Lissette Tolentino
Brian French
Teresa Ober
Kyndra Middleton

INVITED REVIEW PANELS

Yu Bao
Daniel Bolt
Ikkyu Choi
Catherine Close
Steven Culpepper
Christine DeMars
Brian French
Kylie Gorney
Janine Haynes
Corrine Huggins-Manley
Hong Jiao
Matt Johnson
Klint Kanopka
Won-Chan Lee
Chunyan Liu
Susan Lottridge
Kristin Morrison
Aaron Myers
Christopher Ormerod
Leslie Rutkowski
Jim Soland
Shiyu Wang
Jordan Wheeler
John Whitmer
Caroline Wiley

PROPOSAL REVIEWERS

Asiye ŞENGÜL AVŞAR, PhD
Nisha Acharya Julien
Hope Adegoke
Christiana Akande
Sarah Alahmadi
Tony Albano
Çağla Alpayar
Allison Ames
Ernest Amoateng
Benjamin Andrews
Judit Antal
Angel Arias
David Arthur
Nana Amma Asamoah
Nafisa Awwal
Elizabeth Ayers-Wright
Tanushree Banerjee
Erin Banjanovic
Mariana Barragan-Torres
Jonathan Beard
Ahmed Bediwy
Ni Bei
Magdalen Beiting-Parrish
Randy Bennett
Janet Shufor Bih Epse Fofang
Susan Brookhart
Kayla Burt
Heather Buzick
Sandra Liliana Camargo Salamanca
Ian Campbell
Luciana Cancado
Yichong Cao
Chunhua Cao
Víctor H Cervantes
Jianshen (Cassie) Chen
Keyu Chen
Chi-Chen Chen
Xing Chen
Hye-Jeong Choi
Jinah Choi
Jeongwon Choi
Dakota Cintron
Christina Cipriano
Amy Clark
Mary Cochron
Angela Crawford
Justice Dadzie
Laurie Davis
Susan Davis-Becker
Teresa (Tess) Dawber
Juan D'Brot
Mandana delavari
Onur Demirkaya
Yaxin Dong
John Donoghue
Bryan Drost
Lillian Duran
Daniel Edi
George Engelhard
Kerry Englert
Carol Ezzelle
Fen Fan
Denis Federiakin
Richard Feinberg
Zechu Feng
Anthony Fina
Yanyan Fu
Katherine Furgol Castellano
Brian Gane
Tracy Gardner
Kurt Geisinger
Ardeshir Geranpayeh
Melissa Gholson
Lucy Gitiria
Guher Gorgun
Irina Grabovsky
Edith Graf
Yi Gui
Wenjing Guo
Yage Guo
Dongliang Guo
Ahmet Guven
Brian Habing
Syed Abdul Hadi
Brian Harrold
Qiwei He
Surina He
Cristyne Hebert
Amy Hendrickson
Catalina Henríquez
Igor Himelfarb
Kari Hodge
Alexander Hoffman
Cheng Hua
Yue Huang
Mohammad Jahanaray
Hammad Javaid
Jeongmin Ji
Kuan-Yu Jin
Evelyn Johnson
Andrew Jones
Ae Kyong Jung
Juyoung Jung
HyunJoo Jung
Pamela Kaliski
Danielle (Dandan) Kaptur
Yusuf Kara
Hacer Karamese
Leslie Keng
Patrick Kennedy
Eunhee Keum
Marisol Kevelson
Stella Kim
Yun-Kyung Kim
Young Yee Kim
Nana Kim
Yoojoong Kim
Seongeun Kim
Yasmene Kimble
Charalambos (Harry) Kollias
Miryeong Koo
Andrew Krist
Kevin Krost
Huan Kuang
Lavanya Shravan Kumar
Alexander Kwako
Jungwon Kyung
Joni Lakin
Cheryl Lavigne
Youngjun Lee
Hyeryung Lee
Hyunjung Lee
Mina Lee
Arun Balajiee Lekshmi Narayanan
Brian Leventhal
Jennifer Lewis
Shuangting Li
Dongmei Li
Shuhong Li
Jujia Li
Zhen Li
Min Liang
Manqian Liao
Yuan-Ling Liaw
Youn Seon Lim
Sangdon Lim
Huan Liu
Xiaoxiao Liu
Joyce Xinle Liu
Yikai Lu
Ru Lu
Max Lu
Benjamin Lugu
Jinwen Luo
Susan Lyons
Meng Lyu
Wanjing (Anya) Ma
Ye Ma
Hotaka Maeda
Henry Makinde
Mike Maksimchuk
Jaime Malatesta

PROPOSAL REVIEWERS

Kaiwen Man
Scott Marion
Zachary Mayne
Fernando Mena
Qi Meng
Jing Miao
Nicolas Mireles
Moses Mohamed
Mubarak Mojoyinola
Sebastian Moncaleano-Wallrich
William Muntean
Aaron Myers
Mirai Nagasawa
Sebastian Nastuta
Alexander Naumann
Kate Nolan
Steven Nydick
Patrick Obot
Christoper Ocheni
Francis O'Donnell
Valerie Ofori Aboah
Hyeonjoo Oh
Lucy Okam
Comfort Omonkhodion
Timothy O'Neil
Daniel Oyeniran
Jose Palma
Hyemin Park
Junhee Park
Thanos Patelis
Richard Patz
Michael Peabody
Katie Pedley
Lance Piantaggini
Sonya Powers
Xuelan Qiu
Yuxi Qiu
Yale Quan
Victoria Quirk
Mehdi Rajeb
Aileen Reid
He Ren
Kelly Rewley
Michael Rodriguez
Sneha Roy
Andrew Runge
Godwin Sabboh
Shahnaz Safitri
Fusun Sahin
Fusun Sahin
Fariha Hayat Salman
Edgar Sanchez
Purwoko Haryadi Santoso
Merve Sarac
Paulius Satkus
Edynn Sato
Ayfer Sayin
Madeline Schellman
Amy Schmidt
Bob Schwartz
Khem Sedhai
Deepak Sharma
Emily Shaw
Benjamin Shear
Qian Shen
Gozde Sirganci
William Skorupski
Shonai Someshwar
Dan Song
Dorota Staniewska
Sanford Student
Youmi Suk
Tianying Sun
Jeneve Swaby
Cheng Tang
Stephen Tavares
Melinda Taylor
Caitlin Tenison
Danielle Thomas
Zewei Tian
Lissette Tolentino
Mark Tolliver
Anne Traynor
Kelli Treadwell
Nick Trout
Colt Turner
Jon Twing
Stephanie Underhill
Montserrat Valdivia Medinaceli
Dustin Van Orman
Alisha Wackerle-Hollman
Cole Walsh
Siyu Wan
Aijun Wang
Bowen Wang
Xi Wang
Tao Wang
Qi Wang
Jonathan Weeks
Natasha Williams
Stefanie Wind
Tong Wu
Yi-Fang Wu
Zebing Wu
Yi-Chen Wu
Adam Wyse
Daihui Xiao
Yangmeng Xu
Rujun Xu
Yue Xu
Mingfeng Xue
Yiyao Yang
Yiting Yao
Hanwook Yoo
Haimiao Yuan
Peida Zhan
Yuxiao Zhang
Xiuyuan Zhang
Yifan Zhang
Liru Zhang
Yu Zhao
Jialu Zhao
Xinchang Zhou
Hao Zhou
Sizheng Zhu
Tongtong Zou

TRAINING SESSION REVIEWERS

Asiye ŞENGÜL AVŞAR, PhD
Sarah Alahmadi
Allison Ames
Angel Arias
Erin Banjanovic
Jonathan Beard
Ni Bei
Ian Campbell
Luciana Cancado
Jeongwon Choi
Mary Cochron
Daniel Edi
Richard Feinberg
Anthony Fina
Yanyan Fu
Katherine Furgol Castellano
Lucy Gitiria
Yi Gui
Cristyne Hebert
Alexander Hoffman
Cheng Hua
Yue Huang
Jeongmin Ji
Kuan-Yu Jin
Andrew Jones
Ae Kyong Jung
Hacer Karamese
Patrick Kennedy
Marisol Kevelson
Seongeun Kim
Andrew Krist
Huan Kuang
Cheryl Lavigne
Hyunjung Lee
Mina Lee
Youngjun Lee
Arun Balajiee Lekshmi Narayanan
Youn Seon Lim
Ru Lu
Susan Lyons
Henry Makinde
Mike Maksimchuk
Jaime Malatesta
Zachary Mayne
Qi Meng
Jing Miao
Sebastian Moncaleano-Wallrich
Steven Nydick
Hyeonjoo Oh
Yuxi Qiu
Aileen Reid
Kelly Rewley
Andrew Runge
Fusun Sahin
Fusun Sahin
Fariha Hayat Salman
Edgar Sanchez
Paulius Satkus
Bob Schwartz
Khem Sedhai
Deepak Sharma
Gozde Sirganci
Sanford Student
Youmi Suk
Melinda Taylor
Kelli Treadwell
Nick Trout
Montserrat Valdivia Medinaceli
Alisha Wackerle-Hollman
Siyu Wan
Bowen Wang
Yi-Chen Wu
Yi-Fang Wu
Zebing Wu
Yangmeng Xu
Mingfeng Xue
Peida Zhan
Xiuyuan Zhang
Hao Zhou

SCHEDULE AT-A-GLANCE

WEDNESDAY, APRIL 8

Start Time	End Time	Room	Session Title
8:00 AM	10:00 AM	Silver Lake B	Using LLMs for Scalable, Auditable Qualitative Coding & Scoring in Google Sheets
8:00 AM	10:00 AM	Roosevelt A	Introduction to Diagnostic Classification Modeling with R and Stan
8:00 AM	12:00 PM	Hollywood Ballroom II	Bayesian Networks in the Age of AI (A Tribute to Robert Mislevy)
8:00 AM	12:00 PM	Wilshire Grand Ballroom I	Demystify Amazon Web Services (AWS): Cloud Computing and Artificial Intelligence in Psychometric Applications
8:00 AM	12:00 PM	Wilshire Grand Ballroom III	Integrating Generative AI into R Workflows: From APIs to Shiny Apps
8:00 AM	12:00 PM	Silver Lake A	Writing an AI-native dissertation
8:00 AM	12:00 PM	Executive Board Room II	NPCDTools: An R Package for Small-Scale Cognitive Diagnosis and Q-Matrix Theory
8:00 AM	12:00 PM	Hancock Park West	Intersectional Construct Validity: From Factor Structure Exploration to Measurement Invariance Testing
8:00 AM	12:00 PM	Hancock Part East	Organizational Leadership for Measurement Experts
8:00 AM	12:00 PM	Roosevelt B	An Introduction to the Generalized Kernel Equating Framework with Applications in R
8:00 AM	12:00 PM	Hollywood Ballroom I	Applying Data Mining in Multi-Agent Systems for Test Fraud Detection
10:30 AM	12:30 PM	Roosevelt A	Advanced Diagnostic Classification Modeling with R and Stan
10:30 AM	12:30 PM	Silver Lake B	Reliably Connected: Strengthening Your Professional Network
1:00 PM	3:00 PM	Hollywood Ballroom II	Computational Aspects of Psychometric Methods with R (CRC Press, 2023): Practical Tools, Online Resources, and ShinyItemAnalysis
1:00 PM	5:00 PM	Wilshire Grand Ballroom I	Beyond the Score: Creating Rich Feedback with Digital Assessment Data and AI
1:00 PM	5:00 PM	Hancock Park West	Regression Discontinuity Analysis Leveraged by Measurement Models in R Shiny
1:00 PM	5:00 PM	Hollywood Ballroom I	Shedding Light in the Black Box: Understanding What an Item Actually Measures
1:00 PM	5:00 PM	Silver Lake A	Using NIMBLE and iGraph for Diagnostic Classification Bayesian Network Modeling
1:00 PM	5:00 PM	Roosevelt A	Cognitive Diagnosis Modeling: A General Framework Approach and Its Implementation in R
1:00 PM	5:00 PM	Hancock Part East	Longitudinal Diagnostic Classification Models
1:00 PM	5:00 PM	Silver Lake B	Reproducible Educational Reporting using R, RStudio and Quarto

SCHEDULE AT-A-GLANCE

WEDNESDAY, APRIL 8

Start Time	End Time	Room	Session Title
1:00 PM	5:00 PM	Roosevelt B	Vertical Scaling: Hands-on Practice and Evaluation of IRT Methods
1:00 PM	5:00 PM	Wilshire Grand Ballroom III	Evaluating Standard Setting Workshops: Frameworks, Rasch Analyses, and Defensible Reporting

THURSDAY, APRIL 9

Start Time	End Time	Room	Session Title
7:45 AM	9:15 AM	Hollywood Ballroom II	Advancing Diagnostic Models for Fair and Insightful Educational Action (Diagnostic Measurement SIGIMIE Session)
7:45 AM	9:15 AM	Roosevelt A	Methodological Research with Process Data
7:45 AM	9:15 AM	K-Town	Test Taking Behaviors and Response Time Research
7:45 AM	9:15 AM	Majestic	Bayesian Approaches to CAT Research
7:45 AM	9:15 AM	Silver Lake A	Item Parameter Prediction and Difficulty Modeling
7:45 AM	9:15 AM	Boyle Heights	Quantifying and Mitigating Measurement Error
7:45 AM	9:15 AM	Roosevelt B	Remembering Jim Popham
7:45 AM	9:15 AM	Westwood	Three Years After GPT-4: How Has AI Changed Assessments? How Will It?
7:45 AM	9:15 AM	Ladera Heights	Competitor Collaboration: Encouraging Measurement and Assessment Innovation
9:45 AM	11:15 AM	Roosevelt A	Automated Scoring Engine Training: Addressing Real World Constraints
9:45 AM	11:15 AM	Hollywood Ballroom II	Normative Update of A Large-Scale Assessment Of Cognitive Ability: Lessons Learned
9:45 AM	10:45 AM	Hancock Park	Individual eBoards: AI, Technology, and CAT
9:45 AM	11:15 AM	Ladera Heights	Test Security Research
9:45 AM	11:15 AM	Boyle Heights	Using Simulation Studies to Advance Measurement Research
9:45 AM	11:15 AM	Majestic	Test-Taking Behaviors, Strategies, and Interventions
9:45 AM	11:15 AM	Silver Lake A	Classroom Assessment: Fairness and Equity Research
9:45 AM	11:15 AM	K-Town	Evaluating the Effectiveness of AI
9:45 AM	11:15 AM	Westwood	In Memory of Neil Dorans: A Life Dedicated to Equity and Fairness in Assessment
9:45 AM	11:15 AM	Silver Lake B	Moving Measurement Forward, Around, Underneath, Within, and other Prepositions
11:30 AM	1:00 PM	Hollywood Ballroom II	Using AI for Aligning Standards

SCHEDULE AT-A-GLANCE

THURSDAY, APRIL 9			
Start Time	**End Time**	**Room**	**Session Title**
11:30 AM	12:45 PM	Roosevelt A	AI in Medical Assessment: Innovations in Content, Credibility, and Classification
11:30 AM	12:45 PM	Majestic	Validity and Consequential Evidence: Moving Measurement Forward for Alternate Assessments
11:30 AM	12:45 PM	Westwood	Apples and Oranges? Comparing NAEP and State Trends Through Policies and Pandemics
11:30 AM	12:45 PM	Ladera Heights	Innovations in culturally and linguistically responsive measure development
11:30 AM	12:45 PM	Silver Lake A	Data Monitoring: Empowering You to Be a Watch Dog and a Wizard
11:30 AM	12:30 PM	Hancock Park	Individual eBoards: AI and Automated Scoring
11:30 AM	12:45 PM	Boyle Heights	Deconstructing AI Outputs for Bias in the Workplace
11:30 AM	12:45 PM	Roosevelt B	The Design Dialog: How AI Amplifies and Needs Assessment Design Frameworks
11:30 AM	12:45 PM	Silver Lake B	Now what? Leveraging Use Cases for Assessment and Accountability Design and Data
11:30 AM	1:00 PM	K-Town	NCMEntoring: Fireside Chat (Invited Session: Membership Committee)
1:45 PM	3:15 PM	Westwood	From Guessing to Carelessness: Examinee Effort and Its Impact in Assessment Contexts
1:45 PM	3:15 PM	K-Town	Assessing Invariance of Automated Scoring Models
1:45 PM	3:15 PM	Silver Lake A	Estimation and Scoring
1:45 PM	3:15 PM	Boyle Heights	IRT Analyses with Person and Item Covariates
1:45 PM	3:15 PM	Silver Lake B	Modeling Approaches for DCM
1:45 PM	3:15 PM	Ladera Heights	Methodological Issues in Growth Modeling
1:45 PM	3:15 PM	Majestic	Frameworks and Considerations for Dimensionality Assessment
1:45 PM	3:15 PM	Roosevelt A	AI Ethics, Policy, and Practice in Training Measurement Professionals: A Panel Discussion (Educators of Measurement SIGIMIE Session)
1:45 PM	3:15 PM	Roosevelt B	Assessment Needs: Can We Build Balanced Systems of Assessment to Address Them? (Classroom Assessment Committee)
1:45 PM	2:45 PM	Hancock Park	Innovation Demonstrations
3:45 PM	5:15 PM	Roosevelt B	From Topic Models to LLMs: AI-Driven Applications in Educational Measurement
3:45 PM	5:15 PM	Silver Lake A	Advancing Process Data Analysis in Both High and Low Stake Assessments
3:45 PM	4:45 PM	Hancock Park	Graduate Student eBoards: Process data, CAT, technology, and large-scale assessment
3:45 PM	5:15 PM	K-Town	Applications of AI in Psychometrics and Assessment

SCHEDULE AT-A-GLANCE

THURSDAY, APRIL 9

Start Time	End Time	Room	Session Title
3:45 PM	5:15 PM	Majestic	Practical Issues and Innovative Solutions in Standard Setting
3:45 PM	5:15 PM	Silver Lake B	International Large Scale Assessment Research
3:45 PM	5:15 PM	Ladera Heights	Methodological Issues in DCM
3:45 PM	5:15 PM	Westwood	Classifications and Risk Assessment
3:45 PM	5:15 PM	Boyle Heights	Methodological Considerations for DIF Detection
3:45 PM	5:15 PM	Roosevelt A	Understanding Assessment Contexts and Social-Emotional Competencies
3:45 PM	5:15 PM	Hollywood Ballroom II	Collateral Resources to Support Use of the Revised Joint Standards

FRIDAY, APRIL 10

Start Time	End Time	Room	Session Title
9:45 AM	11:15 AM	Roosevelt B	Advancing Conversation-Based Assessment with Large Language Models
9:45 AM	11:15 AM	Hollywood Ballroom II	From Generation to Calibration: Leveraging AI for Item Development, Piloting, and Scoring
9:45 AM	10:45 AM	Hancock Park	Individual eBoards: DIF and Dimensionality
9:45 AM	11:15 AM	Silver Lake A	Integrating AI into Assessment and Psychometric Practice
9:45 AM	11:15 AM	Roosevelt A	Estimation with Complex Distributions and Data Structures
9:45 AM	11:15 AM	Boyle Heights	DIF Research Applications
9:45 AM	11:15 AM	Westwood	Practical issues in CAT
9:45 AM	11:15 AM	Majestic	Leveraging Mixed Methods
9:45 AM	11:15 AM	K-Town	Designing an Assessment System to Support Meaningful Teaching and Learning
9:45 AM	11:15 AM	Wilshire Grand Ballroom III	The Revision of the Testing Standards: Foundations (Invited Joint AERA/ NCME Session)
11:30 AM	12:45 PM	K-Town	Post Launch Measurement: Monitoring Test and Test-Taking Stability
11:30 AM	12:45 PM	Ladera Heights	Evaluating Assessment Practices for Students with Disabilities (Historically Marginalized Groups SIGIMIE Session)
11:30 AM	12:45 PM	Majestic	Redefining Readiness to Recognize Schools' Civic Missions
11:30 AM	12:45 PM	Silver Lake B	Systematically Advancing the Measurement of Inclusion in Schools Worldwide
11:30 AM	12:45 PM	Silver Lake A	Embedding Coherence in Assessment Design and Standard Setting Implementation
11:30 AM	12:45 PM	Westwood	Aggregating Information for Measurement Over Time: A Comparison of 3 Approaches

SCHEDULE AT-A-GLANCE

FRIDAY, APRIL 10			
Start Time	**End Time**	**Room**	**Session Title**
11:30 AM	12:45 PM	Roosevelt A	TACtics: The Future of Technical Advisory Committees for State Testing Programs
11:30 AM	12:30 PM	Hancock Park	Individual eBoards: Methodological Investigations in DCM, G Theory, and IRT
11:30 AM	12:45 PM	Roosevelt B	CANCELLED! When the Data Stops – Implications of Halting Large-Scale Education Data Collections (Committee on Informing Assessment Policy Session)
11:30 AM	12:45 PM	Wilshire Grand Ballroom III	Measurement-Informed Approaches to Evaluate GenAI Outputs (Artificial Intelligence in Measurement and Education SIGIMIE Session)
1:45 PM	3:00 PM	Boyle Heights	Multi-Modal Approaches to Test Security: Patterns, Similarities, and Biometrics (Test Security SIGIME Session)
1:45 PM	3:00 PM	K-Town	Reimagining Large-Scale Assessment—From Measurement to Meaningful Use (Large Scale Assessment SIGIMIE Session)
1:45 PM	3:00 PM	Ladera Heights	Combining Through-Year Scores: Operational Approaches, Challenges, and Practical Implications
1:45 PM	3:00 PM	Westwood	Empirically Evaluating Claims of Instructional Usefulness
1:45 PM	3:00 PM	Silver Lake B	Automated Coding with LLM: Accuracy and Fairness
1:45 PM	3:00 PM	Roosevelt A	Advances in Scoring English Language Learner Items and Responses
1:45 PM	3:00 PM	Majestic	Measurement Challenges in Assessing SEL Competencies in International Longitudinal Studies
1:45 PM	2:45 PM	Hancock Park	Individual eBoard: Practical Issues in Assessment Development and Measurement
1:45 PM	3:15 PM	Roosevelt B	Classroom Assessment: Validity, Implementation, and Score Reporting/ Feedback
1:45 PM	3:00 PM	Wilshire Grand Ballroom III	Featured Session: Threats and Opportunities on the Federal Statistics Landscape
1:45 PM	3:00 PM	Silver Lake A	The Role of Psychometrics in Higher Education Measurement and Assessment
1:45 PM	3:15 PM	Hollywood Ballroom II	NCME Career Achievement Award, Sandip Sinharay: “Reflection on Reporting High-Quality Scores Using Statistics, Measurement, and Professional Judgment”
3:30 PM	5:00 PM	Ladera Heights	Data-Intensive Psychometric Research 2: Return to the Item Response Warehouse
3:30 PM	5:00 PM	Roosevelt A	Advancing Through-Year Assessments: From Technical Evidence to Interpretation of Summative Scores
3:30 PM	5:00 PM	Hollywood Ballroom II	Practical Applications of Artificial Intelligence in the Development of Large-Scale Assessments

SCHEDULE AT-A-GLANCE

FRIDAY, APRIL 10

Start Time	End Time	Room	Session Title
3:30 PM	4:30 PM	Hancock Park	Graduate Student eBoards: IRT and DCM
3:30 PM	5:00 PM	Roosevelt B	Licensure and Certification Research
3:30 PM	5:00 PM	Silver Lake B	Practical Issues in Psychometrics
3:30 PM	5:00 PM	Boyle Heights	Raters and Rating Scales
3:30 PM	5:00 PM	Majestic	Creating and Evaluating Automated Raters
3:30 PM	5:00 PM	Westwood	Featured Session: Battle of the Presidents: Past and Future!
3:30 PM	5:00 PM	Wilshire Grand Ballroom III	Featured Session: NCME Task Force Conversation
3:30 PM	5:00 PM	K-Town	Assessment to Promote Civic Learning: New Volume in the NCME Book Series (Invited Session)

SATURDAY, APRIL 11

Start Time	End Time	Room	Session Title
7:45 AM	9:15 AM	Silver Lake A	Score Comparability Challenges in Applied Assessment Contexts (Contemporary Issues in Scaling, Linking & Equating SIGIMIE Session)
7:45 AM	9:15 AM	Westwood	Innovations in Assessment Development
7:45 AM	9:15 AM	Silver Lake B	Detecting and Assessing DIF
7:45 AM	9:15 AM	Roosevelt B	Visualizing and Interpreting DIF
7:45 AM	9:15 AM	Roosevelt A	Alternate and Personalized Assessment
7:45 AM	9:15 AM	Wilshire Grand Ballroom III	The Revision of the Testing Standards: Operations (Invited Joint AERA/ NCME Session)
7:45 AM	9:15 AM	K-Town	ASA Listening Session with NCME Members: Re-Envisioning the National Center for Education Statistics
8:15 AM	9:15 AM	Hancock Park	Graduate Student eBoards: AI & Machine Learning
9:45 AM	11:15 AM	Boyle Heights	AI Item Difficulty Modeling
9:45 AM	10:45 AM	Hancock Park	Graduate Student eBoards: Validity, Score Reporting, and Accountability
9:45 AM	11:15 AM	Silver Lake B	Scoring Methodologies and Best Practices for Score Reporting
9:45 AM	11:15 AM	K-Town	AI-Assisted Assessment Development
9:45 AM	11:15 AM	Ladera Heights	Automated Scoring Research: Security, Efficiency, and Interpretability
9:45 AM	11:15 AM	Roosevelt A	Equating Research
9:45 AM	11:15 AM	Silver Lake A	Admissions and Higher Education Research
9:45 AM	11:15 AM	Roosevelt B	AI, Machine Learning, & Natural Language Processing Research

SCHEDULE AT-A-GLANCE

SATURDAY, APRIL 11

Start Time	End Time	Room	Session Title
9:45 AM	11:15 AM	Wilshire Grand Ballroom III	The Revision of the Testing Standards: Applications (Invited Joint AERA/ NCME Session)
9:45 AM	11:15 AM	Westwood	In Memory of Robert J. Mislevy: A Legacy of Innovations in Measurement
9:45 AM	11:15 AM	Hollywood Ballroom II	Invited Session: Beyond “National” Measurement: An Introduction to International Assessment Research and Practices
9:45 AM	11:15 AM	Majestic	The Content and Psychometric Partnership: A Panel Discussion (Invited Session)
11:30 AM	12:45 PM	Silver Lake A	Advancing Digital Skills Assessment: Integrating Tests, Performance Data, and Process-Oriented Analyses
11:30 AM	12:45 PM	Boyle Heights	Communicating Technical Results and Concepts to Public Audiences
11:30 AM	12:45 PM	Roosevelt A	Unpacking Item Difficulty Predictor Relationships Across and Within Grades
11:30 AM	1:00 PM	Majestic	Advancing Anti-Racist and Culturally Responsive Assessments Across Contexts
11:30 AM	12:45 PM	K-Town	Validation of Teaching Performance Assessments: Simulation Based Measures of Core Teaching Practices
11:30 AM	12:45 PM	Ladera Heights	AI-Driven Interactive Speaking and Math Assessments: Innovations in Design and Scoring
11:30 AM	1:00 PM	Westwood	Assessing Foundational Competencies: Advancing Educational Measurement through Feedback, Alignment, and Certification
11:30 AM	12:45 PM	Silver Lake B	Applications of Psychometric Methods in Digital Learning Systems
11:30 AM	12:45 PM	Roosevelt B	Generative AI for Implementing Assessment as Learning: From Foundation to Applications
11:30 AM	12:45 PM	Hancock Park	Graduate Student eBoards: DIF, Dimensionality, Growth Modeling, and Equating
11:30 AM	1:00 PM	Hollywood Ballroom II	Navigating NCME Journals: Pathways to Publication (Invited Session: Publications Committee)
11:30 AM	12:45 PM	Wilshire Grand Ballroom III	An NCME Debate Session: The AI Landscape for Industry and Academia (Featured Session Format)
1:45 PM	2:45 PM	Boyle Heights	Validating Literacy Screening Profiles to Identify Risk for Reading Difficulties and Dyslexia
1:45 PM	2:45 PM	Ladera Heights	Development, Validation, and Generalizability Studies of the Compass Classroom Observation Tool
1:45 PM	2:45 PM	Silver Lake B	Innovative Partnerships for Indigenous Language Sustainability and Culturally Grounded Assessments
1:45 PM	2:45 PM	Majestic	Illinois’s Unified Standard Setting: Innovative Strategies to Support Coherence

SCHEDULE AT-A-GLANCE

SATURDAY, APRIL 11

Start Time	End Time	Room	Session Title
1:45 PM	3:15 PM	Hollywood Ballroom II	Psychometric Overlord Auditions Take 2: Prove Me Wrong
1:45 PM	2:45 PM	Hancock Park	Individual eBoards: Scaling, Equating, Linking, and High-Stakes Assessment
1:45 PM	3:15 PM	Roosevelt A	Methodological Issues in Large-Scale Assessment
1:45 PM	3:15 PM	K-Town	Frameworks and Approaches for DCM
1:45 PM	3:15 PM	Westwood	Applications of Network Analysis
1:45 PM	3:15 PM	Roosevelt B	Fairness Concerns with AI
3:30 PM	5:00 PM	Boyle Heights	Text/Speech-based Approaches to Item Parameter Modeling
3:30 PM	5:00 PM	K-Town	From Validity Arguments to Validity Synthesis and Judgment
3:30 PM	5:00 PM	Westwood	Inclusion, Equity, and Fairness in TIMSS and PIRLS
3:30 PM	5:00 PM	Silver Lake B	Reimagining Measurement: History, Method, and Consequences across Higher Education and Statewide Testing (GSIC Session)
3:30 PM	5:00 PM	Hancock Park	Individual eBoards: Classroom & Formative Assessment, Score Reporting, Behavioral Patterns
3:30 PM	5:00 PM	Silver Lake A	Properties of Specific Tests
3:30 PM	5:00 PM	Roosevelt B	Research on Through-Year Assessment
3:30 PM	5:00 PM	Hollywood Ballroom II	Psychometric Issues in Test Development
3:30 PM	5:00 PM	Roosevelt A	Automated Scoring: Combining and Comparing Human and AI Raters
3:30 PM	5:00 PM	Ladera Heights	Sources of Contemporary and Remaining Challenges in Large Scale NGSS Assessment Development

Introduction to Diagnostic Classification Modeling with R and Stan

2-Hour Training Session
8:00 AM – 10:00 AM
Intercontinental Los Angeles Downtown, Floor 7: Roosevelt A

Presenter(s):
Jake Thompson, ATLAS, University of Kansas

Diagnostic classification models (DCMs; also known as cognitive diagnostic models) have gained interest in recent years due to their ability to provide fine-grained actionable feedback while keeping test lengths short. In this workshop, we will cover how to easily estimate and evaluate DCMs with R and Stan. The workshop will include hands-on examples of defining a DCM, estimating the model, and evaluating the fit (e.g., test- and item-level fit, classification accuracy and consistency).

The goal of the workshop is to enable participants to implement DCMs in their own work, and thus is intended for anyone who uses, or would like to use, DCMs for applied or research uses (e.g., psychometricians, faculty, applied researchers, graduate students). Although not necessary, prior experience with R will be helpful. All workshop materials, including slides, examples, and solutions will be available on a workshop website. Participants should have access to a laptop they can bring to the workshop in order to follow along with the examples. Instructions for installing any necessary software will be provided.

Using LLMs for Scalable, Auditable Qualitative Coding & Scoring in Google Sheets

2-Hour Training Session
8:00 AM – 10:00 AM
Intercontinental Los Angeles Downtown, Floor 5: Silver Lake B

Presenter(s):
Max Lu, Harvard University; Rony Rodríguez-Ramírez, Harvard University

This hands on training session will empower participants to run auditable, large scale qualitative coding and scoring with large language models directly in Google Sheets and Docs – no coding required. Using LLM Qual Analyzer (open-source Chrome extension), participants will import qualitative data (e.g., essays or interview transcripts) from their Google Drive, chunk long documents for granular analysis, design rubric aligned prompts, batch run analyses across dozens of files, and log every response to a sheet for a transparent audit trail. We will demonstrate how this versatile tool can be useful for many research questions and empower participants to identify questions they can answer with their own data.

Learning objectives: Build an end to end LLM workflow in Docs and Sheets; craft prompts aligned to coding/scoring needs; evaluate trade-offs in chunking strategies and privacy considerations; batch run and log analyses; replicate runs for IRR computations; compare model settings/providers; export outputs for downstream analysis.

Intended audience: Program evaluators, faculty, and graduate students who need scalable LLM assisted coding/ scoring of unstructured text.

Prerequisites: Google Drive, Chrome Extension read/write/download access

What to bring: a laptop with Chrome.

In session access: we provide pair based temporary API keys (expire after session). Participants may later add their own keys.

An Introduction to the Generalized Kernel Equating Framework with Applications in R

4-Hour Training Session
8:00 AM – 12:00 PM
Intercontinental Los Angeles Downtown, Floor 7: Roosevelt B

Presenter(s):
Jorge Gonzalez, Pontificia Universidad Catolica de Chile; Alina von Davier, Duolingo; Marie Wiberg, Umea University

The aim of test equating is to adjust the score scales on different test forms so that scores can be comparable and used interchangeably. This is extremely important to provide fair assessments to all test takers. The goals of this training session are for attendees to be able to understand the principles of equating, to conduct equating, and to interpret the results of equating in reasonable ways. Emphasis will be given to the new Generalized Kernel Equating (GKE) framework as described in the recently published book “Generalized Kernel Equating using R” written by the instructors (Wiberg, González, von Davier, 2024). Different R packages will be used to illustrate how to perform equating when test scores data are collected under different designs. Traditional equating methods, and both kernel equating and item response theory equating methods under the GKE framework will be illustrated. The main part of the training session is devoted to practical exercises on how to prepare and analyze test score data using different data collection designs and different equating methods. Expected audience should bring their own laptop and includes researchers, graduate students, and practitioners. An introductory statistical background as well as experience in R is recommended but
not required.

Applying Data Mining in Multi-Agent Systems for Test Fraud Detection

4-Hour Training Session
8:00 AM – 12:00 PM
Intercontinental Los Angeles Downtown, Floor 7: Hollywood Ballroom I

Presenter(s):
Kaiwen Man, University of Alabama; Sarah Toton, Caveon; Kylie Gorney, Michigan State University; Jujia Li, University of Alabama; Qipeng Chen, University of Alabama

This session will provide participants with systematic training on applying data mining models in the context of multi-agent systems to detect fraud in diverse test formats, such as computer-based, computer-adaptive, or multistage settings. Using R and/or Python, the session introduces theories and applications of selected supervised and unsupervised methods, including K-Means, Gaussian Finite Mixture, Self-Organizing Maps, K-Nearest Neighbor, Random Forest, Support Vector Machine, and Neural Networks, accompanied by demonstrations. Within a multi-agent framework, test-takers, detectors, and proctoring systems can be modeled as interacting agents, enabling dynamic simulations of fraud strategies (e.g., collusion, item pre-knowledge) and adaptive detection responses.

The session includes lectures, demonstrations, and hands-on activities running commonly used methods, while also highlighting the advantages and limitations of different algorithms and software platforms. It is designed for intermediate to advanced graduate students, researchers, and practitioners interested in both foundational and cutting-edge applications. Some familiarity with R/Python is helpful but not required. Attendees will bring laptops with freely available software installed. By the end, participants will understand how to specify models, run analyses, and integrate data mining approaches within a multi-agent framework to detect aberrant test-taker behaviors, applying these skills directly to their own research and datasets.

Bayesian Networks in the Age of AI (A Tribute to Robert Mislevy)

4-Hour Training Session
8:00 AM – 12:00 PM
Intercontinental Los Angeles Downtown, Floor 7: Hollywood Ballroom II

Presenter(s):
Duanli Yan; Diego Zapata-Rivera, ETS; Russell Almond, Florida State University

With a special tribute to Robert Mislevy who created this annual NCME tutorial in 2002, we present the applications of Bayesian networks in the age of AI.
In the current digital and AI age, information from all direct
ions 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 applications of Bayesian networks including learning progression (using Netica).

Demystify Amazon Web Services (AWS): Cloud Computing and Artificial Intelligence in Psychometric Applications

4-Hour Training Session
8:00 AM – 12:00 PM
Intercontinental Los Angeles Downtown, Floor 5: Wilshire Grand Ballroom I

Presenter(s):
Ye Ma, Amazon Web Services; Vinita Talreja; Mingqin Zhang; Huijuan Meng, Amazon Web Services

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 (ML)/Large Language Models (LLMs)/Generative AI (GenAI) based research and analysis and (2) storing results in the database on the cloud. Participants do not need to have AWS experience. Upon completion, they will be able to streamline ML/LLM analysis along with storing the results in a cloud database with the use of AWS technology. This is a heavy 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.

Integrating Generative AI into R Workflows: From APIs to Shiny Apps

4-Hour Training Session
8:00 AM – 12:00 PM
Intercontinental Los Angeles Downtown, Floor 5 : Wilshire Grand Ballroom III

Presenter(s):
Christopher Runyon, National Board of Medical Examiners

This hands-on workshop is designed for educational measurement professionals with foundational knowledge in psychometrics or statistics who seek to integrate generative AI (GenAI) into R-based workflows. Participants will learn to apply GenAI tools to core assessment tasks while maintaining rigorous measurement standards. The session emphasizes architectural foundations of large language models (LLMs), enabling attendees to make informed decisions about AI integration in various settings.

Learning objectives include understanding key LLM features, applying prompt engineering principles, implementing API interactions, designing multi-agentic systems, and building GenAI-powered Shiny applications.

This session is appropriate for graduate students, psychometricians, assessment developers, and data scientists working in educational measurement or certification contexts. No prior AI experience is required, but working familiarity with R is strongly recommended as basic R instruction will not be provided.
Attendees must bring a laptop with R (v4.4+), RStudio, internet access, and a GenAI API key (free/paid options provided in prep materials). A pre-workshop session will be offered to assist with setup and troubleshooting.
All workshop materials will be available via a public GitHub repository.

Intersectional Construct Validity: From Factor Structure Exploration to Measurement Invariance Testing

4-Hour Training Session
8:00 AM – 12:00 PM
Intercontinental Los Angeles Downtown, Floor 5: Hancock Park West

Presenter(s):
Dakota Cintron, Claremont Graduate University

Ensuring construct validity across diverse populations requires attention not just to single-group measurement invariance but also to differences in factor structures and measurement properties across intersectional groups (e.g., gender × race). This workshop introduces participants to intersectional construct validity testing, explicating exploratory and confirmatory approaches in this context. We begin by discussing how exploratory structural methods and clustering techniques can help identify differences in factor structures across intersectional groups. We then cover confirmatory approaches, including evaluating configural invariance across intersectional groups, followed by demonstrations of the alignment method and moderated nonlinear factor analysis as tools for measurement invariance testing in intersectional contexts.

The workshop is offered in two parts: Part 1 focuses on theory, motivation, and conceptual distinctions between exploratory and confirmatory approaches to intersectional construct validity. Part 2 is a hands-on session applying these methods using R. No prior experience with factor analysis or measurement invariance testing is required for Part 1. Participants attending Part 2 should bring laptops with R and the specified packages (lavaan, openmx) pre-installed.

NPCDTools: An R Package for Small-Scale Cognitive Diagnosis and Q-Matrix Theory

4-Hour Training Session
8:00 AM – 12:00 PM
Intercontinental Los Angeles Downtown, Floor 6: Executive Board Room II

Presenter(s):
Chia-Yi Chiu, Teachers College, Columbia University; Yi-Fang Wu, Cambium Assessment, Inc.; Yu Wang

The training sessions provide theoretically sound and practically useful methods of cognitive diagnosis (CD) with a focus on implementations in small-scale educational settings like classrooms. The sessions provide an introduction to the development of the nonparametric classification methods for small education programs as well as the construction of the Q-matrix, including the properties of complete Q-matrices, Q-matrix validation and estimation, and conditions related to model and Q-matrix identifiability. The goal of the training sessions is to familiarize participants with recent innovations in CD and to provide hands-on experience with the NPCDTools R package and Shiny web apps. The training sessions are of interest to anyone who wishes to use or research CD in small-scale educational settings. Basic knowledge in CD and prior exposure to R would be helpful, but not absolutely required.

Organizational Leadership for Measurement Experts

4-Hour Training Session
8:00 AM – 12:00 PM
Intercontinental Los Angeles Downtown, Floor 5 : Hancock Part East

Presenter(s):
Richard Patz, University of California, Berkeley; Jennifer Dunn, The College Board; Michael Rodriguez, University of Minnesota; Suzanne Tsacoumis, HumRRO

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.

Writing an AI-native dissertation

4-Hour Training Session
8:00 AM – 12:00 PM
Intercontinental Los Angeles Downtown, Floor 5: Silver Lake A

Presenter(s):
Damian Betebenner, Center for Assessment

This half-day training shows how to build an AI-native dissertation that treats AI as a collaborator in analysis, writing, and dissemination. Using an application framework that unifies Quarto (.qmd), an analytics R package, machine-readable APIs, and MCP endpoints, participants will compile a thesis-ready PDF (via their institution's thesis.cls) and publish an interactive website with a chatbot that lets readers explore methods, results, and code.

Learning objectives: (1) design a reproducible repository that cleanly separates text, data, code, and outputs; (2) generate publication-quality artifacts in PDF/HTML from shared sources; (3) expose selected results via REST endpoints and MCP tools so other AI agents can reuse them responsibly.

Intended audience: graduate students, advisors, methodologists, and research software engineers in educational measurement and adjacent fields.

Prerequisites: basic familiarity with R and Git; Quarto experience helpful but not required.

What to bring/install: a laptop with R (≥4.5), Quarto (≥1.6), Git, and RStudio/VS Code. Optional: Node.js (LTS) for web-app extras. A starter repository and small datasets will be provided. We will also discuss institutional policy considerations and reproducible-research ethics when integrating AI into scholarly work.

Advanced Diagnostic Classification Modeling with R and Stan

2-Hour Training Session
10:30 AM – 12:30 PM
Intercontinental Los Angeles Downtown, Floor 7: Roosevelt A

Presenter(s):
Jake Thompson, ATLAS, University of Kansas

Diagnostic classification models (DCMs; also known as cognitive diagnostic models) have gained interest in recent years due to their ability to provide fine-grained actionable feedback while keeping test lengths short. In this workshop, we will cover how to easily customize, estimate, and evaluate DCMs with R and Stan. The workshop will include hands-on examples of specifying customized DCMs and testing alternate specifications (e.g., attribute hierarchies).

The goal of the workshop is to enable participants to implement DCMs in their own work and thus is intended for anyone who uses DCMs for applied or research uses (e.g., psychometricians, faculty, applied researchers, graduate students). Although not necessary, prior experience with R will be helpful. We do assume participants have a basic understanding of DCMs; if not, we recommend completing an introductory workshop first. All workshop materials, including slides, examples, and solutions will be available on a workshop website. Participants should have access to a laptop they can bring to the workshop in order to follow along with the examples. Instructions for installing any necessary software will be provided.

Reliably Connected: Strengthening Your Professional Network

2-Hour Training Session
10:30 AM – 12:30 PM
Intercontinental Los Angeles Downtown, Floor 5 : Silver Lake B

Presenter(s):
Brian Leventhal, James Madison University; Jonathan Henriques, James Madison University; Jada Willse, James Madison University

Networking is essential for professional growth, yet many students and early-career professionals in educational measurement may be unsure how to begin or sustain meaningful connections. This interactive two-hour session offers a structured, inclusive, and reflective approach to networking in academic and applied settings, grounded in NCME's Foundational Competencies in communication and collaboration.
The session will explore how to build professional relationships at the NCME conference and beyond. Through personal stories, a visual network map, and guided discussion, facilitators will highlight how service roles, internships, and mentoring relationships can support long-term network building.

Participants will engage in scenario-based activities, practice crafting and delivering an elevator pitch, and reflect on their current networks. They will have the opportunity to set short- and long-term networking goals tailored to their professional aspirations.

This session is especially relevant for students and early-career professionals, but open to all who want to strengthen their networking skills. No prerequisites or software are required. Attendees will leave with greater confidence when networking and a personalized plan for building meaningful connections that can support their growth in the field of educational measurement.

Computational Aspects of Psychometric Methods with R (CRC Press, 2023): Practical Tools, Online Resources, and ShinyItemAnalysis

2-Hour Training Session
1:00 PM – 3:00 PM
Intercontinental Los Angeles Downtown, Floor 7: Hollywood Ballroom II

Presenter(s):
Patrícia Martinková, Czech Academy of Sciences

This workshop introduces participants to practical computational tools for modern psychometric methods using R, drawing on material from the recent book Computational Aspects of Psychometric Methods with R (Martinkova & Hladka, 2023, CRC Press/Chapman & Hall). The session is designed for researchers, practitioners, and graduate students who seek a hands-on understanding of how psychometric models can be implemented using open-source software. Participants will explore computational strategies such as parameter estimation, model evaluation, visualization, and data simulation, while also receiving practical tips for using the book and accompanying R code in research and teaching.

The workshop will feature a guided exploration of the online materials that support the book, highlighting resources that facilitate independent learning. In addition, participants will gain hands-on experience with the ShinyItemAnalysis app, an interactive platform for learning and teaching psychometric methods and item response modeling.

By the end of the session, attendees will be able to: (1) implement common psychometric models in R with reproducible code, (2) leverage online resources to extend their learning, and (3) use ShinyItemAnalysis to explore psychometric models interactively.

Attendees should bring laptops with R, RStudio, and required packages installed; setup instructions will be provided in advance.

Beyond the Score: Creating Rich Feedback with Digital Assessment Data and AI

4-Hour Training Session
1:00 PM – 5:00 PM
Intercontinental Los Angeles Downtown, Floor 5: Wilshire Grand Ballroom I

Presenter(s):
Hongwen Guo, ETS Research Institute; Matthew Johnson, ETS Research Institute; Luis Saldivia, ETS; Michelle Worthington, ETS; Jeremy Lee, ETS

This half-day training session is a practical and interactive experience aligned with the "Moving Measurement Forward" theme. We will demonstrate how a human-centered AI (HAI) framework can leverage digitally-based assessment data to generate rich, meaningful feedback for teaching and learning. The session will move beyond traditional scoring by showing how response data, process data, and AI models create data-driven insights.

Learning Objectives: Participants will explore a large-scale digital assessment platform, understand the research and development behind AI-driven feedback, and experience an interactive teacher training module firsthand.

Intended Audience: Measurement professionals, educational researchers, assessment & platform developers, and graduate students.

Prerequisites: No prior knowledge of AI or data science is required for attendees.

Materials: Attendees should bring a laptop to participate in hands-on activities. No software needs to be installed for the activities; instructors will provide all necessary links and resources. The session will be a balanced blend of instruction, hands-on activities, and discussion.

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

4-Hour Training Session
1:00 PM – 5:00 PM
Intercontinental Los Angeles Downtown, Floor 7: Roosevelt A

Presenter(s):
Jimmy de la Torre, The University of Hong Kong; Sangbeak Ye, Florida Atlantic University

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.

Evaluating Standard Setting Workshops: Frameworks, Rasch Analyses, and Defensible Reporting

4-Hour Training Session
1:00 PM – 5:00 PM
Intercontinental Los Angeles Downtown, Floor 5: Wilshire Grand Ballroom III

Presenter(s):
Charalambos (Harry) Kollias, Polytomous Limited

Standard setting workshops determine the cut scores that guide high-stakes decisions in education, credentialing, and licensure. Yet reporting of such workshops often lacks systematic evaluation, leaving results open to challenge. This session introduces an expanded evaluation framework for standard setting workshops and demonstrates practical methods for collecting and analyzing evidence to support defensible reporting.
Building on Cizek & Earnest's (2016) framework, recent extensions incorporate Rasch-based analyses (Kollias, 2023) and subgroup evaluation panels (Kanistra, in print, expected 2026). Through a combination of conceptual grounding, group activities, and hands-on software demonstrations, participants will learn how to evaluate procedural, internal, and external validity evidence, conduct interparticipant consistency and decision accuracy analyses, and embed results in transparent validation arguments.

The session is designed for psychometricians, assessment professionals, and test developers. Familiarity with methods such as Angoff or Bookmark is useful but not required. Participants should bring laptops; instructions for installing freely available software will be provided in advance.

Learning objectives: Apply the expanded framework, conduct analyses, and strengthen defensible reporting.

Longitudinal Diagnostic Classification Models

4-Hour Training Session
1:00 PM – 5:00 PM
Intercontinental Los Angeles Downtown, Floor 5: Hancock Part East

Presenter(s):
Matthew Madison, University of Georgia; Madeline Schellman, Pearson

Diagnostic classification models (DCMs) are psychometrics tools that focus on providing actionable feedback in the form of examinee attribute classifications. Longitudinal DCMs have been developed and applied as psychometric tools for analyzing diagnostic assessments administered over multiple occasions. Different from traditional psychometric frameworks for growth, which typically provide continuous and norm-referenced growth estimates, longitudinal DCMs provide categorical growth estimates with criterion-referenced interpretations. This workshop focuses on longitudinal DCMs and their application as the psychometric foundation for categorical growth models. After completing this workshop, participants will understand the structure of longitudinal DCMs, be able to estimate longitudinal DCMs using a newly developed R package, and interpret software output.

This session is appropriate for graduate students, researchers, and practitioners at the emerging or experienced level. Participants are expected to have a basic knowledge of DCMs and psychometrics to enroll. This session presents both conceptual and technical content and also provides hands-on experience for participants to apply what they learn. Content will mostly be delivered through lecture, and content will be reinforced using hands-on activities. Instructor will encourage audience participation through questions and allow time for discussions among participants and the instructor.

Regression Discontinuity Analysis Leveraged by Measurement Models in R Shiny

4-Hour Training Session
1:00 PM – 5:00 PM
Intercontinental Los Angeles Downtown, Floor 5 : Hancock Park West

Presenter(s):
Ji Seung Yang, University of Maryland; Yang Liu, University of Maryland; Muwon Kwon, University of Maryland; Youngjin Han, University of Maryland; Youjin Sung, University of Maryland

Regression discontinuity (RD) analysis is a method used to estimate the effect of a specific intervention when random assignment is infeasible. While conventional RD analysis relies on observed variables, newer approaches combine RD with measurement models, offering important advantages. It allows researchers to better understand how treatment effects vary across individuals, even if they share the same observed score, and draw causal conclusions beyond the cutoff score without reconducting the study. This approach works for both simple and complex data structures (e.g., across schools) and handles various data types, including test scores, surveys, or categorical outcomes.

This training session is designed to make a novel methodological approach in RD analysis accessible to a broad audience, including both methodological researchers interested in causal inference and educational measurement, and program evaluation researchers with limited experience in quantitative methods and programming. The course provides both a theoretical foundation and practical guidance for implementing the analysis using R Shiny. After presenting a high-level overview of the RD analysis, we will offer a hands-on workshop, where participants will follow step-by-step instructions to conduct the analysis using a user-friendly R Shiny application. Participants are encouraged to bring a laptop to fully engage in the practical session.

Reproducible Educational Reporting using R, RStudio and Quarto

4-Hour Training Session
1:00 PM – 5:00 PM
Intercontinental Los Angeles Downtown, Floor 5: Silver Lake B

Presenter(s):
Brian Harrold, University of South Carolina; Erik Whitfield; Nathan Dadey, The National Center for the Improvement of Educational Assessment; Damian Betebenner, Center for Assessment

This half-day session introduces foundations of reproducible analysis and reporting for state assessment and accountability data. The goal is to equip those working with these data, especially within state education agencies, to develop reproducible, modular workflows for generating high-quality, policy-relevant analyses, visualizations, and reports using open-source tools: R, RStudio and Quarto. These tools integrate cleaning, summarizing, modeling, and documentation, enabling participants to transition from ad-hoc analyses to automated workflows that they can reuse annually and share across teams.
Learning objectives include:

1. Understanding the value of reproducible reporting and key considerations.
2. Gaining familiarity with R, RStudio, and Quarto as a unified workflow.
3. Observing demonstrations of existing packages (cohortED and bueller).
4. Building a simple Quarto report to answer a policy-relevant question.
5. Identifying next steps to increase reproducibility within participants' organizations.

The intended audience includes education analysts, psychometricians, and applied researchers working with K-12 assessment data wanting to improve efficiency and reproducibility. Prior experience with R is required. Participants must bring a laptop with R, RStudio, and Quarto installed. Pre-session setup instructions, sample data, and templates will be provided. A toy dataset will be available, and experienced attendees are encouraged to bring their own student-level datasets.

Shedding Light in the Black Box: Understanding What an Item Actually Measures

4-Hour Training Session
1:00 PM – 5:00 PM
Intercontinental Los Angeles Downtown, Floor 7: Hollywood Ballroom I

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

While instrumentalization is neither the first nor the last step of educational assessment, it might be the step where things most commonly go astray. This is perhaps why measurement and assessment professionals without experience as content development professionals (CDPs) are often asked (or volunteer) to evaluate or give feedback on items. Psychometricians, state assessment office personnel, project managers and others sometimes try to chip in, working outside of their disciplinary training and content expertise.

This workshop focuses on basic four protocols and techniques of CDP work. Unpacking standards, recognizing critical issues in items, radical empathy and item alignment examination are key tools for understanding items, how they function and recognizing their level of quality and potential to contribute to validity.

Because non-CDPs are often in a position to offer feedback on items, learning these tools can help them to contribute to item and test quality. They can better understand the issues and requirements around item alignment, fairness and validity from the disciplinary perspective of those responsible for developing high quality items for standard-based assessment. Experienced CDPs who do not know RTD may also find these elements from the Rigorous Test Development framework to be useful additions to their own practice.

Using NIMBLE and iGraph for Diagnostic Classification Bayesian Network Modeling

4-Hour Training Session
1:00 PM – 5:00 PM
Intercontinental Los Angeles Downtown, Floor 5 : Silver Lake A

Presenter(s):
Richard Golden, University of Texas at Dallas; Athul Sudheesh, University of Texas at Dallas

This workshop provides a practical introduction to implementing Bayesian Scoring Engines, a core component of Evidence-Centered Design (ECD), for modern psychometric assessment. Using the flexible NIMBLE and iGraph packages in R, we will demonstrate how this unified framework can be applied to models from the Item Response Theory (IRT), Cognitive Diagnostic Model (CDM), and Structural Equation Model (SEM) families. Participants will gain hands-on experience with an example CDM Bayesian scoring engine development workflow, from model specification to model evaluation. Key learning objectives include: (1) implementing Bayesian scoring engines (e.g., CDMs) using NIMBLE's familiar BUGS/JAGS-like syntax, (2) specifying competency and evidence models using the graphical modeling syntax of iGraph (e.g., "Q" matrix specification), (3) estimating parameters of interest, (4) making psychometric inference, and (5) evaluating model quality using simulation-based experiments. The session is designed for graduate students and researchers interested in flexible and extensible psychometric modeling. No prior experience with specific models is required, but attendees should have intermediate programming experience in R and some familiarity with psychometrics or mathematical statistics. Attendees are strongly encouraged to bring a laptop with current versions of R/RStudio installed; instructions for additional package installation will be provided.

Vertical Scaling: Hands-on Practice and Evaluation of IRT Methods

4-Hour Training Session
1:00 PM – 5:00 PM
Intercontinental Los Angeles Downtown, Floor 7: Roosevelt B

Presenter(s):
Hyeonjoo Oh, Riverside Insights; Hanwook Yoo, Ascend Learning; Tim Moses, Buros Center for Testing

This training demonstrates the empirical application of vertical scaling approaches developed for newly launched large-scale assessments. Vertical scales are a current interest in K-12 testing programs for supporting the measuring and tracking of student growth across grade levels. This session will cover the following topics: (1) conceptual and technical foundations of vertical scaling, including the definitions of student growth, data collection design, and scaling methods, (2) step-by-step guides and hands-on exercises for conducting item response theory (IRT) vertical scaling using both linear and non-linear methods to convert student ability estimates into scaled scores, (3) a hands-on exercise focused on evaluating scaling outcomes and comparing scaled scores, and (4) discussion of practical considerations and future research directions in vertical scaling. Through the training session, participants will gain hands-on experience with empirical vertical scaling and develop an understanding of the challenges and limitations psychometricians encounter during new test development. This training is targeted to advanced graduate students or early-career measurement professionals interested in applying vertical scaling methods. While prior knowledge of IRT-based equating and scaling is recommended, it is not required. Participants should bring their own laptops to engage in the hands-on activities.

NCME Board of Directors Meeting

Invited Session
4:00 PM – 7:00 PM
Intercontinental Los Angeles Downtown, Floor 6: Executive Board Room II

The NCME board meeting is open to all.

Archives & History Committee
Meeting
7:30 AM – 8:30 AM
Intercontinental Los Angeles Downtown, Floor 6: Royal

This committee is responsible for documenting and archiving the history of NCME, and the people, organizations, and trends (e.g., policies, legislation, litigation, societal changes) that have influenced, and have been influenced by, NCME.

NCMEntoring Program
Meeting
7:30 AM – 5:00 PM
Intercontinental Los Angeles Downtown, Floor 6: Roxy

Launched at the 2016 annual meeting, the NCMEntoring Program aims to support the transition of graduate student members and recent graduate members from their graduate programs to professional careers. Early professionals (mentees) are paired with members (mentors) experienced in NCME-related fields: psychometrics, assessment, certification, evaluation, and other aspects of educational measurement.

This experience offers mentees the opportunity to explore possible career paths and/or research interests and for mentors to support the development of potential colleagues and contribute to the field. Each year, over 100 NCME members participate in the NCMEntoring Program and participant feedback has been positive. The Program hopes to cultivate long-term relationships between mentors and mentees.

Advancing Diagnostic Models for Fair and Insightful Educational Action (Diagnostic Measurement SIGIMIE Session)
Coordinated Poster Session
7:45 AM – 9:15 AM
Intercontinental Los Angeles Downtown, Floor 7: Hollywood Ballroom II

This coordinated poster session highlights seven emerging innovations in diagnostic measurement that address practice-oriented challenges in the use of cognitive diagnostic models (CDMs) across real-world educational and assessment contexts. Building on the established foundation, these contributions collectively extend the field through both methodological advancements and applications that generate meaningful, actionable insights. Four methodological studies introduce a marginal item-fit test extending the Diagnostic Facet Status Model to detect Q-matrix misspecification in small-sample contexts, propose the polytomous general nonparametric classification method to flexibly accommodates polytomous attributes and responses, revisit interpretations of mastery probabilities, positioning them as indicators of latent proficiency rather than purely classification certainty, and integrate response time with accuracy data to distinguish misconceptions from disengaged responses. On the applied side, three studies use CDMs to guide instructional design for scientific explanation skills among preservice teachers, assess college students' generative AI literacy to support responsible and effective engagement with emerging technologies, and demonstrate how formative and interim assessments, analyzed in conjunction with Bayesian networks, can identify student performance gaps and inform targeted instruction. Together, these studies demonstrate how diagnostic measurement can bridge psychometric rigor, educational practice, and policy priorities to promote equitable and actionable insights.

Presentations:

1. **Detecting Small Sample Q-Matrix Misspecification in The Diagnostic Facet Status Model**
   Yale Quan, University of Washington; Chun Wang, University of Washington

2. **The Polytomous General Nonparametric Classification Method**
   Hyunjee Oh, Teachers College, Columbia University; Chia-Yi Chiu, Teachers College, Columbia University

3. **Mastery Probability in Cognitive Diagnosis: A Revisit**
   Yiming Chen, University of Minnesota; Nana Kim, University of Minnesota; Wenchao Ma, University of Minnesota

4. **Exploring response time and responses for misconception detection in CDM**
   Tamlyn Lahoud, University of Georgia; Shiyu Wang, University of Georgia

5. **Enhancing Preservice Teachers' Ability to Explain Phenomena Scientifically through Profile-Based Instruction Using the GDINA Model**
   Ivy Mejia, University of Philippines National Institute for Science and Mathematics

6. **Assessing College Students' Generative AI Literacy Using Diagnostic Classification Models**
   Yu Bao, James Madison University; Kobena Eshun, James Madison University; Suyu Wang, James Madison University

7. **Leveraging Assessment Data to Address Educational Performance Gaps**
   Jean Hampel, HMH; Yon Soo Suh, NWEA within HMH

Bayesian Approaches to CAT Research

Individual Paper Session
7:45 AM – 9:15 AM
Intercontinental Los Angeles Downtown, Floor 6: Majestic

Chair:
Leah Feuerstahler (Fordham University)

Discussant:
Leah Feuerstahler (Fordham University)

Presentations:

1. **Bayesian Adaptive Testing for Nominal Responses**
   Presenter(s):
   Luping Niu; Seung Choi, The University of Texas at Austin

   This study extends fully Bayesian computerized adaptive testing (CAT) to the nominal response model (NRM), which models unordered, multiple-category responses. We examine estimation accuracy, test efficiency, and stopping rules under Bayesian NRM-CAT, comparing with conventional approaches. Results demonstrate advantages and practical considerations for implementing Bayesian NRM in adaptive testing.

2. **The effect of informative priors on the test length of a CAT**
   Presenter(s):
   Zachary Mayne, IXL; Michael Peabody, IXL Learning; Yu Zhao

   Bayesian methods suggest that implementing informative priors in a CAT should reduce test length. Simulations under ideal conditions show that this is indeed the case. Real data and simulations using a real item bank illustrate a more complex picture. This study proposes to disentangle the factors that belie conventional wisdom.

3. **Prior Specification for Ability Estimation in Bayesian Adaptive Testing**
   Ae Kyong Jung, University of Iowa; Jonathan Templin, University of Iowa

   In Bayesian adaptive testing, priors for ability estimation are typically specified as normal distributions parameterized by the previous stage's posterior mean and variance. Since this Gaussian approximation is imprecise with limited item information, we propose two alternatives: kernel density approximation and synthetic response generation from the posterior predictive distribution.

4. **Enhancing Ability Estimation Accuracy for Automatically Generated Items in Computerized Adaptive Testing**
   Stella Kim, University of North Carolina Charlotte; Won-Chan Lee, University of Iowa

   This study evaluates the accuracy of an item-parameter adjustment method to improve accuracy of ability estimation for automatically generated items in the context of computerized adaptive testing.

5. **Optimizing Prior Distributions for Dynamic Measurement**
   Erin Banjanovic; Ted Daisher

   Accurately specifying the prior distribution to reflect elapsed time is essential in Bayesian longitudinal measurement. This study evaluates and compares six approaches to specifying the prior distribution in such a model. It concludes by recommending a model capable of tracking student performance over time.

Competitor Collaboration: Encouraging Measurement and Assessment Innovation

Organized Discussion
7:45 AM – 9:15 AM
Intercontinental Los Angeles Downtown, Floor 5: Ladera Heights

We believe that together, we can advance the field of educational measurement and assessment. The purpose of this organized discussion is to bring together five organizations in the K-12 space that, through a Gates-funded initiative, are sharing ideas to innovate assessments, particularly through the lenses of a) AI, b) the connection between classroom and large-scale assessment, and c) iterative approaches to assessment design. We will discuss what we've learned from each other through in-person convenings, how we share ideas across organizations despite (or because of) our overlapping goals, and what we're hoping to achieve in the future. After each organization frames their work and after a guided discussion, we will engage the audience through a question and answer session. The intention is that attendees will leave the discussion with new ideas on how to quickly move forward in the field of measurement with colleagues within and across organizations.

Chair:
Lauren Deters (Khan Academy)

Discussant:
Susan Lyons (Lyons Assessment Consulting)

Presenter(s):
Lauren Deters, Khan Academy; Lanette Trowery, Achievement Network; Steven Ritter, Carnegie Learning; Ourania Rotou, New Meridian Corporation; Angela Bahng, Gates Foundation

Item Parameter Prediction and Difficulty Modeling

Individual Paper Session
7:45 AM – 9:15 AM
Intercontinental Los Angeles Downtown, Floor 5 : Silver Lake A

Chair:
Jiawei Xiong (Curriculum Associates)

Discussant:
Max Lu (Harvard University)

Presentations:

1. **Passage Difficulty Modeling in Oral Reading Fluency Assessments**
   Yusuf Kara, University of Miami; Kuo Wang, Southern Methodist University; Joanne Joo, Southern Methodist University; Zoltan Szentkiralyi, Southern Methodist University; William Annan, Texas Christian University

   Abstract:
   In oral reading fluency assessments, passages contain rich text information that can inform their levels of difficulty. This study explores the feasibility of using features extracted by text mining and natural language processing to predict passage difficulties, which are estimated by a binomial-lognormal speed and accuracy model.

2. **AI-Powered Prediction of 2PLM Parameters from Item Content**
   Yuxiao Zhang, Purdue University; Yanyan Fu, Graduate Management Admission Council; Kyung (Chris) Han, GMAC

   This study evaluated transformer encoders for predicting two-parameter logistic model (2PLM) item parameters from contents. Using DeBERTa-v3-base, we compared frozen embeddings, parameter-efficient fine-tuning (LoRA), and reasoning-trace augmentation with LLMs. Preliminary results showed that targeted fine-tuning improved prediction, offering a computationally efficient, scalable approach to support and streamline item calibration.

3. **The Effect of AI-Derived Item Difficulty on Person Ability Estimates**
   Andrew Krist; Sonya Powers, Edmentum

   Advances in large language models have enabled AI-based estimation of item difficulty. While potentially accurate, these estimates are often systematically biased. This study examines how such bias propagates to person ability estimates in the context of computer adaptive testing, using both simulation and empirical data to assess resulting estimation error.

4. **Using AI Judges' Paired Comparisons to Estimate Item Difficulty**
   Lanrong Li, MetaMetrics, Inc.; Matthew Gushta, MetaMetrics, Inc.

   We examined the performance of a large language model in estimating item difficulty. Five AI judges were created using the model to make paired comparisons of the difficulty of items from a reading test. The results showed that randomizing within-pair item orders changed item difficulty recovery substantially.

5. **Adapting Bookmark Standard-Setting Method with LLMs for Item Difficulty Prediction**
   Ye Yuan, GMAC; Kyung (Chris) Han, GMAC

   Prior studies on using LLMs to predict item difficulty have focused on absolute difficulty, however, direct continuous estimates from LLMs are often poorly calibrated. This study adapts the Bookmark standard-setting method to LLMs for ordinal difficulty prediction. Its categorical nature matches LLMs strengths and improves practical utility for item-parameter prediction.

Latent Space Item Response Models in Action: Advancing Measurement in Emerging Contexts

Coordinated Paper Session
7:45 AM – 9:15 AM
Intercontinental Los Angeles Downtown, Floor 5 : Silver Lake B

Latent space item response models (LSIRMs) extend traditional item response theory by embedding persons and items in a shared geometric space, where distances capture unobserved conditional dependencies beyond what classical models can represent. This coordinated session, themed as Latent Space Item Response Models in Action: Bridging Measurement and Emerging Contexts, brings together five studies that showcase LSIRM's methodological versatility and applied relevance.

The first three presenters address practical questions of dimensionality selection, propose an extension of LSIRM to item response tree models for modeling response styles, and compare LSIRM with network psychometrics to highlight the complementary perspectives on binary assessment data analysis. Building on these methodological advances, two applied studies demonstrate LSIRM's reach in practice and emerging context: examining gender differences in autism diagnosis alongside autism representations on social media, and evaluating research productivity and collaboration networks in a mobile health training program.

By integrating methodological development with socially relevant applications, this session demonstrates LSIRM's potential as both a rigorous and practical framework. It emphasizes the model's capacity to enhance validity, fairness, and interpretability in modern educational and psychological measurement, and can inspire how psychometric innovations can move measurement forward.

Chair:
Yingshi Huang

Discussant:
Paul De Boeck (The Ohio State University)

Presentations:

1. **Dimension Selection Guideline for Latent Space Item Response Models**
   Yingshi Huang, University of California - Los Angeles

2. **A Latent Space Item Response Tree Model for Modeling Response Styles**
   Nana Kim, University of Minnesota

3. **Network Approaches to Binary Assessment Data: Network Psychometrics vs. LSIRM**
   De Carolis Ludovica, University of Milano-Bicocca

4. **Using LSIRM to Examine Gender Differences and Social Media Portrayals of Autism**
   Ingrid Tien, Centre for Addiction and Mental Health, McCain Centre for Child, Youth, and Family Mental Health

5. **Advancing Program Evaluation through LSIRM: Mapping Scholar Productivity and Collaboration**
   Jinwen Luo, UCLA

Methodological Research with Process Data

Individual Paper Session
7:45 AM – 9:15 AM
Intercontinental Los Angeles Downtown, Floor 7: Roosevelt A

Chair:
Bowen Wang (National Board of Chiropractic Examiners)

Discussant:
Logan Rome (Curriculum Associates)

Presentations:

1. **Latent Poisson Count Models for Action Count Data From Technology-Enhanced Assessments**
   Gregory Arbet, University of Texas at Austin; Hyeon-Ah Kang, University of Texas Austin

   The study presents latent Poisson count models tailored for action count data from technology-enhanced assessments. We refine the Rasch and Conway-Maxwell Poisson models to account for item-type effects, and evaluate their performance through Monte Carlo simulations and applications to real-world data.

2. **A Bi-Factor Model for Latent Trait Estimation Integrating Responses and Process Data**
   Huan Kuang, Florida State University; Jiawei Xiong, Curriculum Associates; Okan Bulut, University of Alberta; Justin Kern, University of Illinois at Urbana-Champaign

   This study introduces a bi-factor IRT model integrating item responses and LSTM-modeled process data. Using data from 257 U.S. examinees in PISA 2012, we compare theta from a traditional IRT model with the bi-factor general factor. The general factor correlated strongly with plausible values, while specific factors captured process-related variance.

3. **Using Functional Outcomes from Response Process Data for Accurate Ability Estimation**
   Yuxuan Li, Teachers College, Columbia University; Youmi Suk, Teachers College, Columbia University

   This study introduces how to improve ability estimation with functional outcomes from response process data. We construct scalar summaries using functional principal component analysis and incorporate these summaries to IRT models for accurate ability estimation. We demonstrate the effectiveness of our approach using a simulation study and a real-data application.

4. **Identifying Response Patterns Using Growth Mixture Modeling in TIMSS 2023**
   Xiaoxiao Liu, University of Alberta; Surina He; Steven Wise, EngagedMeasurement; Ying Cui, University of Alberta

   This study identified three response patterns (fast response, interest-oriented, and stable response) using screen-based response times. It also examined how each pattern was related to test effort and attitudes toward mathematics using TIMSS 2023 data. Results show that math confidence, disorderly behavior and test effort significantly influence stable respondents' performance.

5. **Detecting Non-Effortful Responses Using Residual RTs and GLMM Trees**
   Hyerim Noh; Hyun Sook Yi, Konkuk University

   Validity of test scores depends on examinee engagement, yet low-stakes assessments frequently produce non-effortful responses. Existing detection approaches are hindered by cutoff dependence and computational limitations. This study introduces an RT residual–GLMM tree model, using the 2024 Korean National Digital Literacy Assessment, identifying disengagement patterns while enhancing precision and generalizability.

Quantifying and Mitigating Measurement Error

Individual Paper Session
7:45 AM – 9:15 AM
Intercontinental Los Angeles Downtown, Floor 5 : Boyle Heights

Chair:
Andrés Christiansen (IEA Hamburg)

Discussant:
Michael Peabody (IXL Learning)

Presentations:

1. **The Differential Impact of Careless Responding Patterns on Coefficient Omega Estimates**
   Mohammed Abulela, MetaMetrics, Inc. and University of Minnesota; Kyle Nickodem, University of Minnesota - Twin Cities; Michael Rodriguez, University of Minnesota

   We conducted a simulation study investigating how careless responding (CR) affects categorical omega reliability estimates while varying CR pattern type, prevalence, severity, sample size, and population reliability. We found bias in categorical omega estimates was positively associated with CR prevalence and severity with CR patterns having differential impacts.

2. **Deriving Stratified-Omega Coefficients for Composite Scores using Multivariate Structural Equation Models**
   Walter Vispoel, The University of Iowa; Hyeryung Lee, Oklahoma State University; Tingting Chen, The University of Iowa

   Reliability coefficients for composites formed from subscale scores should be, but rarely are, adjusted for subscale representation and interrelationships. Although Cronbach et al. (1965) developed stratified-alpha coefficients with such adjustments, they have the same shortcomings as conventional alpha coefficients. We use SEMs to produce stratified-omega coefficients that address these limitations.

3. **Quantifying Effects of Item Wording and Measurement Error Using Extended Bifactor Models**
   Walter Vispoel, The University of Iowa; Hyeryung Lee, Oklahoma State University; Tingting Chen, The University of Iowa

   We use extended bifactor models for multi-occasion data from self-report measures to quantify effects of traits, item wordings, and multiple sources of measurement error. Models including all effects provided good fits with total item wording and measurement error effects, respectively, accounting for up to 8.5% and 24.8% of total-score variance.

4. **Modeling Completion Rate in Assessments Using Bayesian Hierarchical Models**
   Paulius Satkus; John Willse, Udemy; Alexandra Lay, Udemy

   Variability in assessment completion threatens validity in low-stakes contexts. Using 95 skill assessments, we model continuation across item positions with Bayesian hierarchical binomial trajectories. Model comparison favors a logarithmic curve—steep early drop, gradual taper. Random intercepts dominate heterogeneity; slopes are similar. Findings suggest front-loaded persistence interventions and a disengagement methodology.

5. **Evaluation of Distractor Model Specifications within Assessment Engineering**
   Sunhyoung Lee, Ascend Learning/ University of Nebraska-Lincoln; James Bovaird, University of Nebraska-Lincoln

   Various types of distractor models within the Assessment Engineering framework hold potential for efficient model-based item development. This study presents an enhanced evaluation of how different types of distractor models affect psychometric properties, such as item difficulty and discrimination, by varying test stakes and item calibration methods.

Remembering Jim Popham
Invited Session
7:45 AM – 9:15 AM
Intercontinental Los Angeles Downtown, Floor 7: Roosevelt B

In this tribute to Jim Popham, his life, work, and impact on the NCME community will be remembered. All are welcome to share reflections in honor and celebration of Jim.
Chair: Ellen Forte (edCount, LLC)

Speakers:
Ellen Forte, edCount, LLC, Derek Briggs, University of Colorado – Boulder, Chad Buckendahl, ACS Ventures, LLC , Gerunda Hughes, Stephen Sireci, University of Massachusetts Amherst

Test Taking Behaviors and Response Time Research
Individual Paper Session
7:45 AM – 9:15 AM
Intercontinental Los Angeles Downtown, Floor 5: K-Town

Chair:
Mina Lee (Cambium Assessment)

Discussant:
Mina Lee (Cambium Assessment)

Presentations:

1. **Comparisons of Response Time-Informed IRT-, MICE-, and Autoencoder-based Imputation Methods across Countries and Languages**
   Surina He; Usama Ali, ETS; Peter van Rijn, Cito

   This study compares IRT-, MICE-, and autoencoder-based methods for imputing planned missing responses in PISA 2022. Using response accuracy and response times across multiple countries and languages, we evaluate imputation accuracy, recovery of item means and IRT parameters, and computational efficiency, providing new insights into scalable methods for large-scale assessments.

2. **Investigating Student Test Quitting Using Response Time Data**
   Bin Tan, University of Alberta; Xiaoxiao Liu, University of Alberta; Okan Bulut, University of Alberta

   This study examines whether students' early quitting of low-stakes, computer-based assessments can be predicted and explained using screen-level response times (RTs). Deep learning and n-gram analyses reveal that quitting can be accurately predicted as early as the 4th-5th screen, with sustained slowness or very fast pacing as the primary predictors.

3. **Analyzing Process Data to Diagnose Item Characteristics: Eye-Tracking Insights**
   Yizhu Gao, University of Georgia; Shiyu Wang, University of Georgia; Yaxuan Yang, University of Georgia

   Using eye-tracking from 70 undergraduates who completed 20 spatial-rotation items, we examined three process indicators—option entropy, section entropy, and scan-path complexity—as complements to psychometrics. These indicators align with item difficulty and discrimination and differentiate efficient verification from exploratory search, offering actionable guidance for item design, interpretability, and validation.

4. **Classifying Test-Taking Behaviors and Navigational Patterns on a High-Stakes Admissions Assessment**
   Yi Yang, ETS; Guangming Ling, ETS

   Using data from a computer-based higher education admissions exam, this study extends prior research by classifying test-taking behaviors and navigational patterns, analyzing their associations with performance, test content, and demographics. Findings advance understanding of navigation in computer-based testing and inform fairness, validity, and design considerations in large-scale assessments.

5. **A Multi-Modal Framework for Modeling Strategy in a Systems Thinking Game-Based Assessment**
   Sizheng Zhu, Roblox; Matt Emery, Roblox; Xinchu Zhao, Roblox; David Laing, Roblox; Erica Snow, Roblox; Jack Buckley, Roblox

   This study introduces a framework for analyzing multi-modal data from a systems thinking game-based assessment. The framework uses a clustering algorithm on extracted features to identify strategic profiles, Hidden Markov Models to map latent behavioral sequences in process data, and topic modeling to decode self-reported strategies from essays.

Three Years After GPT-4: How Has AI Changed Assessments? How Will It?

Organized Discussion
7:45 AM – 9:15 AM
Intercontinental Los Angeles Downtown, Floor 5 : Westwood

In NCME 2023, just a few weeks after the release of GPT-4, we organized a panel discussion on the potential implications of GPT-4 and other advanced AI tools in educational assessment. We held a follow-up discussion one year later at NCME 2024. Now, for NCME 2026 (three years after GPT-4's debut), we propose a new panel discussion to examine how the landscape of assessment has been reshaped by the transformative influence of powerful AI in both industry and academia. Our discussion will center on what has truly worked, what remains hype, and what the future may hold for assessment. We will focus on six key areas: skills and constructs, item and task development, psychometric methodologies, reimagining assessment administration, upskilling psychometric practitioners, and preparing the next generation of measurement professionals. We hope this panel discussion will catalyze the exchange of ideas and insights across the assessment community, guiding us toward a more cohesive understanding of AI's transformative potential in assessment.

Chair:
Jiangang Hao (ETS)

Discussant:
Victoria Yaneva (National Board of Medical Examiners)

Presenter(s):
Lauren Deters, Khan Academy; Lanette Trowery, Achievement Network; Steven Ritter, Carnegie Learning; Ourania Rotou, New Meridian Corporation; Angela Bahng, Gates Foundation

Individual eBoards: AI, Technology, and CAT

Electronic Board Session
9:45 AM – 10:45 AM
Intercontinental Los Angeles Downtown, Floor 5 : Hancock Park

Presentations:

- **RAG-Based Iterative Text Generation for Fine-Tuned CEFR Classification Models in Language Assessment (eBoard 1)**
  Ayfer Sayin, Gazi University

  This study aims to examine AI-based text generation for CEFR classification in Turkish as a foreign language. A RAG-based iterative text generation method is proposed to create training data for fine-tuned models. Results demonstrate promising outcomes for CEFR-based classification, with ongoing project phases exploring broader applications across proficiency levels.

- **AI-Driven Detection of Enemy Items in MCQs: A Preliminary Framework (eBoard 2)**
  Hope Adegoke, University of North Carolina, Greensboro; Henry Makinde; Godwin Sabboh, University of North Carolina - Greensboro

  Enemy items compromise test validity when one question reveals the answer to another. We present a preliminary AI-based framework that combines embeddings and NLI to detect those overlaps. Using OpenBookQA and Q3 local dependence statistics as weak supervision, we explore the feasibility of a scalable method that identifies item dependencies.

- **Beyond Item Generation: Large Language Models (LLMs) for Item Evaluation and Selection (eBoard 3)**
  Meltem Ozcan, University of Southern California

  This research investigates large language models (LLMs) as tools for item evaluation and selection in test development. Using experiments across multiple constructs, models, and prompt strategies, expert–LLM alignment was analyzed with multilevel models. Findings reveal variation across LLMs and conditions, highlighting opportunities and limitations for psychometric applications in education.

- **Test Items of a Feather Cluster Together: Intelligent Blueprinting with NLP (eBoard 4)**
  Josiah Hunsberger, James Madison University; Aquia Richburg, National Commission on Certification of Physician Assistants; Marcus Walker, National Commission on Certification of Physician Assistants; Yanlin Jiang, NCCPA; Andrew Dallas, NCCPA

  Traditional test blueprints often overlook the multidimensional, overlapping nature of exam content. Using transformer embeddings, UMAP, and HDBSCAN, we clustered a large medical licensing exam. Findings show cluster-based blueprinting provides a pseudo-multidimensional perspective that complements SME judgment, supports adaptive and fixed-form assembly, and refines test design and item bank management.

- **Designing AI-Enabled Formative Assessment Tools for Scientific Argumentation (eBoard 5)**
  Teresa Ober, ETS; Nichole Jusino del valle, ETS Research Institute; Yi Song, ETS Research Institute; Field Watts, ETS; Lei Liu, ETS

  This study examines challenges and opportunities in designing AI-enabled formative assessment tools for scientific argumentation. After iterative development and refinement, two AI components—a feedback generator and a conversational agent—were discussed and evaluated for usability and instructional alignment. Findings highlight design principles for transparency, adaptability, and integrated feedback systems.

- **Hyperparameter Optimization for Neural Network-Based IRT in Constrained Settings (eBoard 6)**
  Seong Eun Hong, Center for Applied Linguistics

  This study investigates hyperparameter optimization in neural network-based IRT estimation. Using simulation with varying item pool sizes, iterations, and tuning strategies, results show fine-tuning improves performance under constrained settings, offering practical guidelines for implementing neural networks in educational assessment with limited item banks and test administrations.

- **Comparative Analysis of NLP-Based Models for Readability Prediction (eBoard 7)**
  Wei Xu, ISC2; Donna Butterbaugh, ISC2

  Readability is an essential element of reading comprehension. In this study, we compared the different readability formulas and models (BERT, GCN, RAIT, Gradient Boosting Regressor). We found the ensemble approach that incorporates both NLP and traditional readability metrics have proved to be more effective.

- **What AI Sees Before Data: Construct Validity in Psychological Tests (eBoard 8)**
  Youngmi Cho, Riverside Insights; Hyeonjoo Oh, Riverside Insights

  This study examined whether large language model (LLM)–derived embeddings recover the latent structure observed in human responses. A six-domain noncognitive instrument and a Big Five benchmark were used to evaluate AI-assisted, pre-data construct validation. Results indicate that embeddings mirror empirical structure and provide an early signal of construct clarity.

- **Improving Construct Distinction Through Prompt Engineering and Embedding-Based Diagnostics (eBoard 9)**
  Youngmi Cho, Riverside Insights; Hyeonjoo Oh, Riverside Insights

  This study applies large language model embeddings and prompt-engineered item refinement to a multidomain assessment. Iterative semantic diagnostics guided targeted revisions that reduced cross-domain overlap while maintaining domain cohesion. Findings demonstrate the promise of embedding-guided methods as pre-data tools for strengthening construct distinction and assisting test development.

- **MxML: Using AI agent to review the latest ML/AI literature in measurement (eBoard 10)**
  Yi Zheng, Arizona State University; Sijia Huang, Indiana University Bloomington; Steven Nydick, Duolingo; Susu Zhang, University of Illinois at Urbana-Champaign; Abhave Abhilash, Arizona State University

  Systematic literature reviews are important but time-consuming components of research. Leveraging recent developments of Generative AIs, we develop an AI agent for extracting themes and providing insights from the MxML literature. The feasibility and efficacy of the developed AI agent are examined by comparing its results against human coding.

- **Detecting Aberrant Responses within Testlet Items: A Machine Learning Approach (eBoard 11)**
  Wei Xu, ISC2; Donna Butterbaugh, ISC2

  Detecting aberrant response behaviors help ensure test validity. In this study, we evaluate the performance of an ensemble approach that combines Gradient Boosting with person-fit measures on detecting aberrant responses for testlet-based items. Model performance across different testlet effects is evaluated using false-positive rate and true-positive rate.

- **Machine Learning Ability Estimator for Multidimensional Item Response Theory (eBoard 12)**
  Jingyi Guo, University of Notre Dame; Cheng Liu, University of Notre Dame

  In Multidimensional CAT, MAP estimation often suffers from prior-induced bias, especially at the early stage with limited data. To address this, we developed a new estimator combining MAP information with machine learning. For K=3, our method outperforms MAP in both RMSE and bias, notably for individuals with extreme abilities.

- **Unrestricted p-Optimality Method for Item Pool Design and Extension (eBoard 13)**
  Jyun-Hong Chen, National Cheng Kung University; Hsiu-Yi Chao, Soochow University

  This study introduces the unrestricted p-optimality method for item pool design, using an item-centered approach to eliminate the arbitrary bins of conventional methods. A Simulation study showed the method creates a nearly 40% smaller blueprint with comparable measurement precision, demonstrating a more efficient and cost-effective framework for developing item pools.

- **Repercussions of Switching from MAP to MLE Scoring in Computerized Adaptive Testing (eBoard 14)**
  Adam Wyse, Renaissance; Edison Choe, Renaissance Learning

  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.

- **Bias and Fairness in Two-Stage Multistage Testing Under 3PL Models (eBoard 15)**
  Whitney Thomas, University of Nebraska - Lincoln; Sarah Hammami, University of Nebraska-Lincoln; Jordan Wheeler, University of Nebraska - Lincoln; Tim Moses, Buros Center for Testing

  This simulation study examines ability estimate bias in two-stage multistage testing using the three-parameter logistic item response theory model. Results demonstrate substantial bias reduction for correctly routed examinees but concerning bias reversal patterns for misrouted students. Findings have important implications for test design and fairness.

- **Good Kitty, Bad Bank? Rescoring Miscalibrated CATs Improves Accuracy (eBoard 16)**
  Xingyao Xiao, Stanford University; Benjamin Domingue, Stanford University; Michael Frank, Psychology, Stanford University; Klint Kanopka, New York University

  We study how adaptive test scores are affected when item banks are small, shifted in difficulty, or used across different populations. Simulations show rescoring with improved calibration recovers much accuracy, though poor item selection still limits performance. Results point to rescoring and better selection rules as practical fixes.

- **How Many Items Are Enough? Minimum Item Pool Size Recommendations for CAT (eBoard 17)**
  Yu Zhao; Michael Peabody, IXL Learning; Zachary Mayne, IXL

  This study explores minimum item pool requirements for computerized adaptive tests when population information is unavailable. Simulation findings offer a general rule of thumb for item needs when developing an item pool, and provide practical guidance on ensuring precise ability estimation, especially when resources are limited.

- **Measurement Precision and Test Content with Linear-on-the-Fly Testing (LOFT) (eBoard 18)**
  Susan Embretson, Georgia Institute of Technology

  Linear-on-the-fly test (LOFT) design has many advantages as compared to traditional fixed-length tests in mitigating item exposure. This study examines the impact of varying LOFT designs on trait estimation and test content as compared to traditional parallel forms. The results indicate that structured LOFT designs provide the best results.

- **Calibrating Gamification in a Digital Learning Environment (eBoard 19)**
  Fusun Sahin, Curriculum Associates; William Spagnola, Curriculum Associates

  Gamification in digital learning platforms can bolster student engagement. However, precisely calibrating a motivation system can be challenging. We illustrated implementing a hypothetical system for earning Experience Points (XP) and reaching levels for i-Ready Personalized Instruction. Results explored the students' attainment of XP and levels, which impacted existing rules.

Automated Scoring Engine Training: Addressing Real World Constraints

Coordinated Paper Session
9:45 AM – 11:15 AM
Intercontinental Los Angeles Downtown, Floor 7: Roosevelt A

This coordinated paper session addresses real-world constraints in training automated scoring engines for educational assessment. Five papers present empirical research and methodological innovations aimed at improving the efficiency, scalability, and fairness of automated scoring systems. The first paper evaluates specialized encoder-only language models for long-form essay scoring, highlighting their effectiveness with extended responses. The second paper investigates how optimizing training data through rater behavior can reduce sample size requirements while maintaining model quality, with implications for new or revised assessments. The third paper systematically examines the impact of training sample size on model performance using advanced encoder-only architectures and adaptation techniques. The fourth paper demonstrates that strategic hyperparameter optimization can substantially reduce the amount of training data needed for operational deployment of transformer-based scoring models. The fifth paper explores the use of chain-of-thought prompting and preference optimization in open-source generative models to enhance rubric alignment and scoring consistency. Collectively, these studies offer actionable insights for assessment programs deploying automated scoring in resource-constrained environments, with direct implications for policy and operational practice.

Chair:
Edward Wolfe (Iowa Testing Programs / University of Iowa)

Discussant:
Andrew Lan (University of Massachusetts, Amherst)

Presentations:

1. **Train before you Test: Strategies for Optimising Automarker Training Data**
   Mark Brenchley, Cambridge Assessment; Trevor Breakspear, Cambridge Assessment; Abhirup Chakravarty, Cambridge Assessment; Hannah Bouteba, Cambridge Assessment; Theerdha Sajimon, Cambridge Assessment; Ian Lewin, Cambridge Assessment; Yan Huang, Cambridge Assessment

2. **Exploration of Language Models for Automated Scoring of Long Essays**
   Hong Jiao, University of Maryland; Haowei Hua, Princeton University; Hanna Choi, University of Maryland - College Park; Sydney Peters, University of Maryland, College Park; Xinyi Wang, University of Maryland

3. **Fine-Tuned Thinking LLMs for Scoring and Feedback**
   Yilong Wang, Cambium Assessment; Alexander Kwako, Cambium Assessment; Christopher Ormerod

4. **Systematic Hyperparameter Exploration for Minimal Training Data**
5. **in Transformer-based Scoring Models**
   Michael Hemenway, Pearson; Martha Bellows, Pearson; Justin Barber, Pearson

6. **Minimal Data, Maximum Performance: A Parameter-Efficient Framework for Automated Scoring**
   Alexander Kwako, Cambium Assessment; Gitit Kehat, Cambium Assessment; Christopher Ormerod; Mackenzie Young

Classroom Assessment: Fairness and Equity Research

Individual Paper Session
9:45 AM – 11:15 AM
Intercontinental Los Angeles Downtown, Floor 5: Silver Lake A

Chair:
Ivy Mejia (University of Philippines National Institute for Science and Mathematics)

Discussant:
Michael Hardy (Stanford University)

Presentations:

1. **Implementing Multilingual Multimodal Classroom Assessment Activities: Evidence from Elementary School Science**
   Keira Ballantyne, Center for Applied Linguistics; Amy Simpson-Burden, Center for Applied Linguistics; Brittany York, Center for Applied Linguistics; Silvia McDonald, North Carolina Dept of Public Instruction

   Multilingual students in K-12 settings are learning and being assessed in content areas while still in the process of learning the language of the assessment. Our research project explores how classroom educators might use multilingual multimodal formative assessment activities to gain a clearer picture of multilingual elementary students' scientific sense-making.

2. **Advancing Equity in Chilean Teacher Assessment Literacy Through Scenario-Based Measurement**
   Megan Welsh; Valeria Zunino, University of California, Davis; Maria Santelices, Pontificia Universidad Católica de Chile

   This study reports on the development and validation of the LEEC, a scenario-based measure of Chilean teacher candidates' assessment literacy. Using a justice-oriented validity framework, we examined content alignment, psychometric properties, and equity considerations, offering insights for reframing the conceptualization of assessment literacy to include cultural responsiveness and justice.

3. **Using Formative Assessment Practices to Promote the Learning of Students with Disabilities**
   Tara Kintz, Michigan Assessment Consortium; Kristy Walters, Corunna Public Schools; Edward Roeber, Michigan Assessment Consortium; Sheryl Lazarus

   This session examines how formative assessment practices can equitably support students with disabilities. Drawing on research and professional learning through Michigan's FAME program, findings highlight ways educators plan, adapt, and implement formative assessment to address learner variability. Implications for policy, practice, and socioculturally-responsive measurement are discussed.

4. **Initial Sound Errors in Kindergarten Students' Spelling: Patterns and Implications**
   Dukjae Lee, University of Virginia; Carlin Conner, University of Virginia; Tisha Hayes, University of Virginia; Kerry Shea, University of Virginia; Laura Darcy, University of Virginia; Emily Solari, University of Virginia

   This study examined spelling mistakes made by kindergarten students in identifying the initial sound of one-syllable words with short vowels. Common mistakes included confusing similar-sounding consonants and omitting parts of digraphs or blends. These findings have implications for analyzing spelling responses of kindergarten students at high risk for reading difficulties.

5. **Investigating Multidimensional & Unified Approaches to Assess Conceptions of the Structure of Matter**
   Linda Morell, University of California, Berkeley; Mark Wilson, University of California, Berkeley

   For the study, we examined a middle school science assessment previously developed for diagnostic purposes to investigate its potential for instructional planning purposes. We found that by designing the four diagnostic dimensions in a way that aligns with a single overall construct, we could achieve good results for both.

Evaluating the Effectiveness of AI

Individual Paper Session
9:45 AM – 11:15 AM
Intercontinental Los Angeles Downtown, Floor 5 : K-Town

Chair:
Erick Sandi

Discussant:
Hongwen Guo (ETS Research Institute)

Presentations:

1. **GenAI Evidence Hub: Applying Measurement Science to AI Research Studies in Assessment**
   John Whitmer, Learning Data Insights; Magdalen Beiting-Parrish, EdAIfy, CUNY Graduate Center; Alexis Andres, Learning Data Insights

   The rapid proliferation of GenerativeAI has created an urgent need for a measurement-informed analysis of the results of these applications. This paper describes results from a structured literature review of GenAI research to create a framework of emergent practices in validity, reliability and fairness analysis for research and development projects

2. **Can LLMs Replicate Student Testing Behaviors? A Cognitive Diagnostic Modeling Approach**
   Xiuxiu Tang, University of Notre Dame; Yikai Lu, University of Minnesota Twin Cities; Ying Cheng, University of Notre Dame

   Large language models (LLMs) show potential as human substitutes in research, yet their capacity to authentically replicate student testing behaviors remains underexplored. Using TIMSS data and cognitive diagnostic modeling (CDM), we benchmark LLMs' performance against human responses and CDM simulations, examining how prompting strategies affect psychometric fidelity and response authenticity.

3. **Evaluating Large Language Models' Performance for Psychometric Simulation Studies in R**
   Mohammed Abulela, MetaMetrics, Inc. and University of Minnesota; Ethan Brown, Fulbright University Vietnam

   We present an evaluation of three leading large language reasoning models to generate R simulation code corresponding to two published Educational Measurement studies, incorporating modern best practices in simulation design and reporting. The results highlight differences between models and emerging best practices for prompt design for psychometric simulations.

4. **Robustness of Embedding Similarity for Content Alignment Across Models and Indices**
   Josiah Hunsberger, James Madison University; Paulius Satkus

   Measurement professionals increasingly use embeddings for NLP tasks. We examine how similarity-indices and encoder choices affect conclusions in the State–NAEP standard alignment context. Across grades, these choices shift distributions and rankings. Hence, valid interpretation hinges on method-aware calibration and agreement across indices, rather than fixed, portable cutoffs.

5. **The Efficacy of AI Personas in Mimicking Human Assessment Response Data**
   Heather Hayes, Western Governors University; Nathan Sundt, Western Governors University

   This study investigates whether AI-generated data can mimic human responses to competency-based educational (CBE) assessments. We incorporate skills, personality, and motivation profiles into this process and compare performance and psychometric properties of assessments between AI and human data. Findings evaluate AI's potential for supplementing human data during CBE assessment development.

In Memory of Neil Dorans: A Life Dedicated to Equity and Fairness in Assessment

Invited Session
9:45 AM – 11:15 AM
Intercontinental Los Angeles Downtown, Floor 5 : Westwood

This presentation will focus on the totality of the contributions of Neil Dorans to educational measurement. Brief discussions will cover each of Neil's primary areas of interest, including differential item functioning, item response theory, score scales, score linking, population invariance, and score equity analysis. The current relevance of Neil's contributions to the field will also be discussed.

Chair:
Randy Bennett (Assessment Innovation Matters)

Discussant:
Sandip Sinharay, ETS

Presentations:

1. **Score Equity Assessment and the Pursuit of Fairness:**
   Continuing Neil Dorans' Legacy

   Jinghua Liu, The National Board of Osteopathic Medical Examiners

   In my talk, I will honor the late Dr. Neil J. Dorans by sharing the Score Equity Assessment (SEA) work I conducted with him and reflecting on his enduring influence on our field. I will briefly introduce SEA as a structured approach for evaluating whether score interpretations and uses are comparably supported across groups and contexts, illustrate its application with examples from our collaborations, and distill practical lessons for assessment programs seeking to embed equity evidence in routine validation. Throughout the presentation, I will highlight Neil's singular gift for connecting research and practice—translating psychometric insights into actionable guidance for testing programs—along with his generous mentorship and support that shaped my career and many others. The session aims to honor Neil's legacy while offering concrete tools and principles attendees can adapt to advance fairness and validity in their own work.

2. **Philosophical Perspectives on Test Fairness**
   Rebecca Zwick, University of California, Santa Barbara

   I'll discuss the chapter Neil and I wrote together for the volume, Fairness in Educational Assessment and Measurement, for which Neil and Linda Cook were editors. For years, Neil and I had talked about coauthoring a technical paper on differential item functioning, involving such questions as how best to define the reference and focal groups and how test-takers could best be matched on proficiency. For the new volume, we discarded those ideas and addressed an entirely different and largely non-technical question: How would prominent philosophers—we landed on Aristotle, Robert Nozick, and John Rawls—regard some key educational measurement issues of our lifetime, involving college admissions criteria, opportunity to learn classroom material, and the determination of scholarship eligibility scores. Given that neither of us were philosophers by training, working on the paper turned out to be a fascinating, frustrating, and sometimes hilarious experience. (What would Aristotle have thought about the SAT?) I'm so glad I had the chance to work with Neil on this unusual project.

3. **Revisiting Dorans' Lingering Linking Issues: An Updated Discussion**
Tim Moses, Buros Center for Testing

At AERA, 2022, Neil Dorans gave a presentation to accept the 2021 Robert L. Linn Distinguished Address Award. I was his Discussant, which was an honor and an opportunity to revisit some of my score linking collaborations with Neil. It was also a great opportunity to consider Dr. Linn's work in more detail. The treatment of score linking in our field has clearly evolved from Linn to earlier and later works by Neil, particularly in terms of how predictions, correlations, reliabilities and group dependencies are regarded. Terminology, approaches and views on best practice are being refined to this day, and Neil was helping me with these refinements until very recently. In this talk, I will summarize some aspects of score linking that Neil considered especially important in his AERA, 2022 presentation and also in our recent exchanges. Neil concluded his AERA, 2022 presentation by stating that his points would need to be made again and again as the years go by, an unending task he would leave to younger generations. I will close with a similar charge that emphasizes the score linking issues that are lingering, periodically updated, and always well worth revisiting.

4. **The Continuing Evolution of Fairness Assessment**
Michael Walker, HumRRO

In the last decade, Neil Dorans focused much of his attention on fairness assessment in a pluralistic society. We see this emphasis in the Dorans and Cook chapter, "The implications of societal changes for fairness assessment," which appeared in the 2016 volume he edited with Linda Cook. Dorans, Syp, and Walker (2022) continued that theme, exploring different perspectives on fairness as well as the challenges to statistical approaches to fairness assessment. He returns to this theme in his forthcoming chapter on the evolving role of conditions in fairness assessment, which I also had the privilege of working with him on. In this talk I will discuss this most recent work. I will also share some personal notes on Neil as provocateur, whose work led to the formation of so many pearls of measurement.

Moving Measurement Forward, Around, Underneath, Within, and other Prepositions

Organized Discussion
9:45 AM – 11:15 AM
Intercontinental Los Angeles Downtown, Floor 5 : Silver Lake B

Fact: None of us are getting any younger (we're already older than when we read the proposal title).
Fact: Stress and life pressures continue to compound.
Fact: We take ourselves so seriously that we don't realize how ridiculous we can be.
Fact-ish: We need to laugh.
A systematic review and meta-analysis show that laughter can greatly reduce cortisol levels (Kramer & Leitao, 2023).
The goal of this session: to laugh at ourselves and the work that stresses us out so much (and delay that death thing).
This session includes SNL-like sketches, comedic commercials, volunteers, and audience participation. We'll satirize applications and uses of Artificial Intelligence; Diversity, Equity, and Inclusion; and Validity and Fairness.

Examples:

- A snake oil peddler will show a few volunteers how the magic salve of AI solves their testing administrations and implementations.
- A plea to ban interracial marriages because of its negative impact on DIF, tracking, and policymaking.
- A demo categorizing people by behaviors better than demographics (e.g., Tik Tok dances, Name that Tune).

If you saw the session in 2021, this is like that. But in person. And in 2026. And, if you haven't, well… Get ready. Buckle up. This ain't your usual NCME session. Giddyup!

Chair:
Pamela Paek

Presenter(s):
Pamela L Paek (Primary Presenter), Chad Buckendahl, ACS Ventures, LLC, Britte Haugan Cheng, Menlo Education Research, Susan Lottridge, Pearson,Ellen Forte, edCount, LLC

Normative Update of A Large-Scale Assessment Of Cognitive Ability: Lessons Learned

Coordinated Poster Session
9:45 AM – 11:15 AM
Intercontinental Los Angeles Downtown, Floor 7: Hollywood Ballroom II

Developing national norms for large-scale assessments is a complex and multifaceted process, particularly in the post-pandemic era. This session presents the norming methodology for the Cognitive Abilities Test, covering key components such as data collection, sampling procedures, weighting strategies, and statistical modeling. It also explores the impact of pandemic on students' performance, emerging trends in the updated norms, and comparisons with the normative update of an achievement test, the Iowa Assessments. Additionally, the session discusses implications for school districts, especially in the context of Gifted and Talented program identification. Designed for those who are not familiar with norm development, this coordinated poster session offers a practical and accessible introduction to norm development, highlighting its significance in educational measurement and decision-making.

Presentations:

1. **What is the CogAT, and Why Do We Need a Normative Update?**
   Joni Lakin, University of Alabama

2. **Data Sampling, Stratification, and Weighting**
   Matthew Naveiras, Riverside Insights; Sharon Frey, Riverside Insights

3. **Investigating the Impacts of the COVID-19 Pandemic on CogAT**
   Sid Sharairi, Riverside Insights; Matthew Naveiras, Riverside Insights

4. **Developing the Post-COVID CogAT Norms**
   Sharon Frey, Riverside Insights; JongPil Kim, Riverside Insights

5. **Overall Trends in CogAT Scores Between 2017 and 2024**
   Onur Demirkaya, Riverside Insights; Matthew Naveiras, Riverside Insights; Sharon Frey, Riverside Insights

6. **Comparison with the Iowa Assessments Normative Update**
   Jing Ma, University of Iowa; Yen Vo, The University of Iowa; Anthony Fina, University of Iowa

7. **Impact of the CogAT Normative Update on the Gifted-Talented Student Selection**
   JongPil Kim, Riverside Insights; Matthew Naveiras, Riverside Insights; Sharon Frey, Riverside Insights; Evelyn Johnson, Riverside Insights

Test Security Research

Individual Paper Session
9:45 AM – 11:15 AM
Intercontinental Los Angeles Downtown, Floor 5: Ladera Heights

Chair:
Kylie Gorney (Michigan State University)

Discussant:
Jon Twing (HumRRO)

Presentations:

1. **Cheating Detection Using Biclustering: Improving Accuracy and Selecting Optimal Cutoffs**
   Hyeryung Lee, Oklahoma State University

   This study proposes Rasch-augmented biclustering for cheating detection. The method reduced false positives while preserving sensitivity, outperforming baseline biclustering. In addition, the elbow point of accumulated p-values provided a practical decision rule for bicluster retention, balancing sensitivity and specificity to strengthen the practical utility of cheating detection.

2. **Enhancing the Use of Response Times in Answer Similarity Analysis**
   Nick Trout, Michigan State University; Kylie Gorney, Michigan State University

   We develop a new version of the ω statistic that incorporates response times and accounts for the particular items on which similar answers are observed. Results suggest that the new version of the statistic controls the Type I error rate and is more powerful, on average, than existing versions.

3. **Real-Time Test Security Monitoring Using Machine Learning Data Drift Detection**
   James Ingrisone, Pearson VUE; Soo Ingrisone, HumRRO

   This study proposes machine learning-based data drift detection for real-time test security monitoring. Using the Evidently framework and cosine similarity analysis on compromised certification exam data, results demonstrate effective anomaly detection, demonstrating the framework's potential for providing testing organizations with immediate breach alerts during active testing.

4. **An Algorithm for Identification of Preknowledge in Speeded Assessments**
   Stuart Barnum, National Board of Osteopathic Medical Examiners; Min Liang, National Board of Osteopathic Medical Examiners (NBOME)

   We develop an algorithm for identifying preknowledge in speeded assessments. Because of rapid guessing and faster responses by examinees with preknowledge, particularly for correct answers, we focus on response times for correct answers. Facilitating application of Bernoulli mixture models and k-means, the algorithm is tested with simulated and real data.

5. **Multi-source Process Data Boosts Item Compromise Detection**
   Qipeng Chen, University of Alabama; Kaiwen Man, University of Alabama; Susu Zhang, University of Illinois at Urbana-Champaign; Jeffrey Harring, University of Maryland

   To leverage fixation count in compromised item detection, the present study proposes a Bayesian change-point model that integrates response accuracy, response time, and fixation counts. Simulations show that incorporating fixation counts improves detection accuracy and parameter recovery.

Test-Taking Behaviors, Strategies, and Interventions

Individual Paper Session
9:45 AM – 11:15 AM
Intercontinental Los Angeles Downtown, Floor 6: Majestic

Chair:
Mina Lee (Cambium Assessment)

Discussant:
Jiangang Hao (ETS)

Presentations:

1. **Decoding AI Tutor Effects for Educational Measurement: Temporal, Multi-Outcome, and Behavioral-Cognitive Analysis**
   Yiyao Yang, Teachers College, Columbia University; Yasemin Gulbahar, Teachers College, Columbia University

   The research proposes a novel framework to examine AI tutor effects on data mining learning using educational data. Temporal interaction patterns forecast performance and trust, multi-outcome analyses reveal trade-offs across learning and perception metrics, and behavioral-cognitive clustering uncovers latent student profiles, informing personalized and adaptive AI-assisted instructional strategies.

2. **How Students' Computational Thinking Strategies and Engagement Guide an Automated Bus**
   Yuan-Ling Liaw, IEA Hamburg; Mojca Rozman, IEA Hamburg

   This study uses ICILS 2023 Automated Bus, where students program navigation and brake systems through diagrams, decision trees, and simulations. Latent profile analysis identifies computational thinking strategies. Results indicate that revision, reflection, and ICT-rich teaching practices are linked to achievement. Both system-level factors and student characteristics shape students' approaches.

3. **Tracing Students' Experimentation Strategies Through Sequential Process Data Mining**
   Yizhu Gao, University of Georgia; Hee-Sun Lee, The Concord Consortium; Tongxin Zhang, Beijing Normal University; Fu Chen, University of Macau; Chang Lu, Shanghai Jiao Tong University; Lijing Wang, Shandong Second Medical University; Zipei Zhu, University of Georgia

   Using 681 simulation logs, we traced how students' experimental strategies changed during a scientific investigation. Sequential process mining revealed four patterns. Multilevel models showed these patterns significantly predicted observation accuracy, evidence quality, and explanation level. These findings suggest that suboptimal strategies may function as steppingstones toward systematic experiment designs.

4. **An Exploratory Structural Equation Modeling Test of a Computer Science Pedagogical Intervention**
   Tom McKlin, The Findings Group; Diley Hernández, Georgia Institute of Technology; Jessica Summers, University of Arizona; Jayma Koval, Georgia Institute of Technology

   The purpose of this study was to use Exploratory Structural Equation Modeling to test the theory of change of an intervention program aimed at broadening computer science education for Latino students. Results highlight the complex relationship between identity identification, belonging, self-efficacy, career persistence, gender, culture and context.

5. **Educators' Knowledge about Measuring Achievement Gaps**
   Joselyn Perez

   Limited research exists on educators' data literacy regarding achievement scores. This study examines educators' understanding of the reliability of school-based achievement scores and factors influencing them. It also assesses changes in their knowledge after interacting with an online demonstration designed to improve their interpretation of achievement data.

Individual Paper Session
9:45 AM – 11:15 AM
Intercontinental Los Angeles Downtown, Floor 5: Boyle Heights

Chair:
Fathima Jaffari (Psychometric Expert at Etec Qiyas)

Discussant:
Kristin Morrison

Presentations:

1. **Using Generative AI for Sequential Data Generation in Monte Carlo Simulation Studies**
   Youmi Suk, Teachers College, Columbia University; Pan Chenguang, Teachers College, Columbia University; Ke Yang, University of Texas at San Antonio

   This study proposes an AI-based simulation framework that leverages generative AI to create realistic synthetic data and incorporate them into Monte Carlo simulation studies. Our framework has five key steps: pre-processing input data, training AI models on input data, assessing synthetic data quality, conducting AI-based simulations, and evaluating simulation results.

2. **Validity of Misfitted Models in Multidimensional Forced-Choice Testing**
   Sirui Wu, University of British Columbia; Amery Wu, University of British Columbia

   This study investigates whether latent trait scores from multidimensional forced-choice tests remain valid when the fitted model does not match the actual decision process. Simulations across three models and varying inter-trait correlations clarify how misfit affects ipsative and normative interpretations, offering guidance on model choice in practical context.

3. **Getting Real with Monte Carlo Simulations**
   Richard Feinberg, NBME; Carolin Strobl, University of Zurich

   Based on an upcoming chapter, this session reviews the evolution of Monte Carlo applications and offers practical guidance for designing, conducting, and evaluating simulations. Our intent is to help attendees think through realistic considerations for their research, such as appropriate research questions, how many replications, and how to communicate results.

4. **A Comparison of Field-Test Item Calibration Methods under Model Misspecification**
   Yiting Yao, Florida State University; Yanyun Yang, Florida State University

   Computerized adaptive testing (CAT) relies on the assumption of unidimensionality for ability estimation and item calibration. This study investigates the joint effects of model misspecification and calibration method (the Stocking A method and fixed-parameter calibration method) on parameter recovery under violation of unidimensionality in CAT.

5. **Variational Estimation Multidimensional Factor-augmented Regularized Latent Regression To Analyze Complex Large-scale Assessment**
   Yijun Cheng; Chun Wang, University of Washington; Gongjun Xu, University of Michigan

   We propose GVEM-MFARLR, a multivariate extension of factor-augmented regularized latent regression for large-scale assessments. Leveraging a Gaussian variational EM algorithm, our method efficiently estimates multidimensional latent traits. Simulations show improved bias and RMSE over traditional DIRE, offering a flexible, scalable approach for analyzing complex assessment data.

Diagnostic Measurement SIGMIE
Meeting
10:00 AM – 11:00 AM
Intercontinental Los Angeles Downtown, Floor 6: Royal

Diagnostic measurement is a well-established subfield of educational and psychological measurement that focuses on the development, modeling, and interpretation of diagnostic assessments.
Since the publication of Diagnostic Measurement: Theory, Methods, and Applications (Rupp, Templin, & Henson, 2010), there has been a wave of research centered around the development and application of diagnostic measurement methodologies. The Diagnostic Measurement SIGMIE was established to connect diagnostic measurement practitioners and researchers around common goals of 1) increasing awareness, accessibility, and professional development opportunities in diagnostic measurement; 2) advancing diagnostic measurement methodology; and 3) creating an interdisciplinary and collaborative community of applied and methodological researchers.

Discord Networking
Meeting
11:15 AM – 12:15 PM
Intercontinental Los Angeles Downtown, Floor 6: Royal

Individual eBoards: AI and Automated Scoring
Electronic Board Session
11:30 AM – 12:30 PM
Intercontinental Los Angeles Downtown, Floor 5 : Hancock Park

Presentations:

- **The Rush to ChatGPT: Are We Overlooking Bias in AI-Generated Educational Assessments? (eBoard 1)**
  Mohammed Abulela, MetaMetrics, Inc. and University of Minnesota; wessam mohamed

  To examine how LLMs' training-data tendencies may yield item-level bias in ChatGPT-generated items, we analyzed responses from 810 undergraduates. Using Mantel–Haenszel, logistic regression, and Lord's χ^2 across sex and study discipline, alongside sensitivity analyses, we detected multiple C-level DIF items, underscoring the need to investigate systematically bias in AI-generated assessments.

- **Automatic Item Generation for a Language Assessment Using Large Language Model Prompting (eBoard 2)**
  Yage Guo; Yamei Wang, Center for Applied Linguistics; Anna Zilberberg, Center for Applied Linguistics

  This study investigates the application of a large language model (LLM) to automatically generate complex, open-ended items for a language proficiency assessment, utilizing three distinct prompting strategies. The findings provide practical guidance on designing prompts that optimize LLM output for complex, domain-specific assessment tasks.

- **Automatic Item Generation Using Multi Agent System (eBoard 3)**
  Henry Makinde

  This study investigates a multi-agent LLM framework for automated item generation (AIG) in educational assessment. Findings will demonstrate that coordinated AI agents can produce sound items comparable to traditionally authored questions while significantly reducing development time (Lee et. al. 2025). Results include quality expert reviews and psychometric properties.

- **Automated Metadata Classification in Item Development Using Deep Learning and Transformer Architectures (eBoard 4)**
  Joe Betts, National Council of State Boards of Nursing; William Muntean, National Council of State Boards of Nursing

  This study explores advanced deep learning and transformer models to automate metadata classification of assessment items, including those generated by GenAI. By expanding model complexity and metadata scope, it aims to improve classification accuracy and reduce reliance on manual SME coding in test development.

- **Can AI Evaluate its Own Test Items? A Study of LLM Self-evaluation (eBoard 5)**
  Xiaochen Xu, University of California, Davis; Tony Albano, University of California - Davis; Maxime Pouokam, University of California, Davis; Leyao Xing, Harvard University

  This study examines whether AI can reliably evaluate AI-generated statistics items by comparing the psychometric properties of human-written items, human-selected AI items, and AI-selected AI items across two undergraduate courses. Using GPT-5 for generation and Gemini-2.5 for evaluation, we test AI's potential to reduce human review burden while maintaining assessment quality.

- **Dynamic Parameter Estimation for AI-Generated Test Items Using Explanatory IRT Approach (eBoard 7)**
  Ahmed Bediwy, The University of Iowa; Jonathan Templin, University of Iowa

  High-quality assessments depend on efficient item development, yet traditional piloting is costly. This study proposes a framework for dynamic parameter estimation of AI-generated math items using explanatory item response models (EIRMs). By utilizing GPT-4.1 and modeling item features, we improve parameter prediction, reduce piloting, and advance automated assessment design.

- **Harnessing LLMs for Automated Item Generation in Social Studies and Science Assessments (eBoard 8)**
  Jinah Choi, Edmentum; Sonya Powers, Edmentum

  This study investigates the feasibility, validity, and efficiency of AI-generated social studies and science items. Disciplinary frameworks and expert input are used to develop AI prompts. AI-generated items are human reviewed. An iterative refinement process is used to explore the strengths and limitations of LLMs currently in item development.

- **Assessing Operational Items in an Automated Item Generation (AIG) Context (eBoard 9)**
  Jonathan Beard, College Board

  AIG produces large pools of test items by systematically varying item features within templates. While this offers advantages for test development efficiency and security, the evaluation of these generated items remains critical. This study proposes a scope and sequence to evaluate test questions that are optimally designed to be isomorphic.

- **Comparison of Machine Learning and Large Language Models on Item Response Prediction (eBoard 10)**
  Huan Liu, Riverside Insights; Evelyn Johnson, Riverside Insights; Hsin-Ro Wei, Riverside Insights; Tong Wu

  This study examines how accurately machine learning (ML) and large language models (LLMs) can predict student responses to field-test items on a social-emotional learning assessment. Findings indicate that non-fine-tuned LLMs perform at a level comparable to traditional ML methods. Fine-tuned LLMs are expected to substantially outperform ML approaches.

- **Using IRT methods to Correct Scale Shrinkage Problem in Automated Scoring (eBoard 11)**
  Ru Lu, Educational Testing Service; Shuhong Li, Educational Testing Service; Venessa Manna

  Automated scoring improves efficiency and consistency in assessments but often suffers from scale shrinkage, where scores are compressed compared to human ratings. This reduces variance and affects fairness and validity. With empirical data, this study evaluates whether using IRT methods can address scale shrinkage in automated scoring.

- **Enhancing Open-responses Analysis for a Post-exam Survey with NLP and LLM Tools (eBoard 12)**
  Yunyi Long, National Board of Osteopathic Medical Examiners; Xia Mao

  This study examined an approach for analyzing open-ended responses from post-exam surveys for a high-stakes medical examination by integrating NLP-based classification with LLM-assisted summarization. The AI-generated summaries closely aligned with human-produced summaries and demonstrated reasonableness. The feasibility of this approach and considerations on confidentiality of using AI-assisted tools are discussed.

- **Comparing Chain-of-Thought Reasoning and Direct Instruction Prompting for Automated Essay Scoring (eBoard 13)**
  Ayfer Sayin, Gazi University; Mark Gierl, University of Alberta

  This study compares two AI-supported automated essay scoring approaches for Turkish essays. Human raters independently scored responses and established adjudicated standards. Two few-shot models were compared: an analytic chain-of-thought scorer producing rationales and an instruction-based holistic scorer. Findings underscore the importance of method selection and provide suggestions for future research.

- **Synthetic Data Generation for Automated Essay Scoring Models (eBoard 15)**
  Sungjin Nam, ACT, Inc

  We examine whether simple lexical features like length can enhance instructions for a large language model and help create a high-quality essay dataset. Our initial results show that the scoring model trained with synthetic essays generated by our optimized instructions performs significantly better than the baseline outputs.

- **Evaluating Creativity with Multimodal LLMs: Prompt and Task Effects on Human-AI Agreement (eBoard 16)**
  Haeju Lee, University of North Carolina Greensboro; Jinmin Chung, Univ. of Iowa; Sungyeun Kim, Incheon National University

  We evaluated creativity through title-writing for two different images using multimodal LLMs (GPT-5, GPT-4o). The effects of model, prompt, temperature, and image on AI-human agreement were analyzed using Spearman and Pearson correlations, RMSE, and multi-faceted Rasch models, highlighting conditions where AI ratings closely match expert judgments.

- **AI Scoring and Feedback Generation on a Formative Situational Judgment Test (eBoard 17)**
  Cole Walsh, Acuity Insights; Rodica Ivan, Acuity Insights; Gabriel Sitarenios, Acuity Insights

  This study presents an AI-powered system for scoring and generating personalized feedback on an open-response situational judgment test (SJT) assessing personal and professional skills. Our LLM-based approach demonstrates scoring reliability comparable to human experts and provides detailed, actionable feedback, offering a scalable solution for integrating formative SJTs into academic programs.

- **Accuracy and Reliability of Generative AI Scoring of Essays (eBoard 18)**
  Lei Wan, College Board; Merve Sarac, College Board; Amy Hendrickson, The College Board; Serena Magrogan, College Board

  We examined the accuracy and reliability of a few large language models' scoring of essays. Additionally, a qualitative analysis was conducted to investigate why sometimes AI scoring differed significantly from human scoring. The results inform the potential of LLMs to perform automatic essay scoring in educational assessments.

- **Scoring Simulated Conversations with Small Training Datasets: Comparing LLMs and Fine-Tuned Models (eBoard 19)**
  Alessia Marigo, Wisconsin University - Madison

  Simulations provide preservice teachers with opportunities to practice complex skills and receive timely feedback. While human scoring these performances is expensive and time-consuming, and many AI-supported scoring methods rely on extensive datasets, the two AI approaches we propose require smaller datasets, making them a viable solution for small, customized studies.

- **Incorporating AI Confidence into Theta Estimation in IRT Models (eBoard 20)**
  Kate Nolan, Amira Learning; Ran Liu, Amira Learning; Qi Qin, Cognia; Kang Xue, Lexia learning

  This paper introduces a method for incorporating AI model confidence into item weighting during IRT scoring. Simulation and empirical analyses demonstrate that confidence-informed modulation of item discrimination improves theta estimation, offering a novel psychometric use of AI metadata and enhancing trust in AI-assisted assessments.

- **Going Natural: Using NLP for Scoring Constructed Response Items (eBoard 21)**
  Blaine Pedersen

  Scoring constructed response (CR) items on educational assessments is often resource-intensive, especially when items are scored by hand. This study employs simple natural language processing techniques to achieve CR scoring efficacy that approaches performance criteria common in the assessment industry, which can save costs and time associated with scoring.

- **Is Using Rationales Rational? (eBoard 22)**
  Daniel McCaffrey, ETS; Jodi Casabianca, BroadMetrics; Matthew Johnson, ETS Research Institute; Mo Zhang, Educational Testing Service

  This study investigates multi-stage prompting of large language models (LLMs) for automated essay scoring, focusing on the validity of LLM-generated rationales. By comparing methods with and without rubric access, we examine construct relevance, alignment with human ratings, and practical implications for fair, transparent, and scalable educational assessment.

- **Scoring Consistency Across and Within Large Language Models (eBoard 23)**
  Mingfeng Xue, University of North Carolina Greensboro; Xingyao Xiao, Stanford University; Yunting Liu, University of California, Berkeley; Mark Wilson, University of California, Berkeley

  Large language models (LLMs) excel in automatic scoring. However, scoring inconsistency can occur within an LLM and across LLMs. We investigate the intra-LLM and inter-LLM consistency in scoring with five LLMs and under different temperatures and propose a voting strategy to address inconsistent scoring.

AI in Medical Assessment: Innovations in Content, Credibility, and Classification

Coordinated Paper Session
11:30 AM – 12:45 PM
Intercontinental Los Angeles Downtown, Floor 7: Roosevelt A

This session explores the transformative role of artificial intelligence in modern medical assessment design and validation. Through four interconnected presentations, we examine how AI tools can enhance exam development, streamline qualitative analysis, improve citation accuracy, and automate cognitive classification.
The first presentation compares human-led and AI-driven thematic analyses of focus group data to support the integration of new constructs, such as professionalism, into certification exams. The second introduces a citation verification pipeline using the NCBI API to correct and validate AI-generated references, addressing a growing concern in academic writing. The third presentation evaluates the quality and efficiency of AI-generated multiple-choice items compared to human-authored content, offering insights into subject matter expert (SME) feedback and editing time. Finally, the fourth presentation showcases BloomBERT, a fine-tuned transformer model that classifies exam questions by cognitive level, demonstrating its potential to automate Bloom's Taxonomy-based analysis in high-stakes testing.

Together, these studies highlight AI's potential to reduce expert workload, improve fairness, and support scalable innovation in assessment. Attendees will gain practical strategies and methodological insights for responsibly integrating AI into exam development workflows.

Presentations:

1. **Charting New Content Territory: AI and Human Theming in Exam Development**
   Susan Hibbard, The American Board of Anesthesiology

2. **Enhancing Citation Accuracy: Leveraging the NCBI API to Verify and Correct AI-Generated References**
   Ting Wang, American Board of Family Medicine

3. **AI-Generated Items: A Comparison with Human Authorship**
   Heath Kincaid, American Board of Obstetrics and Gynecology

4. **AI-Based Cognitive Level Classification Using BloomBERT**
   Ying Du, American Board of Pediatrics; Hope Adegoke, University of North Carolina, Greensboro

Apples and Oranges? Comparing NAEP and State Trends Through Policies and Pandemics

Coordinated Paper Session
11:30 AM – 12:45 PM
Intercontinental Los Angeles Downtown, Floor 5 : Westwood

NAEP and State tests share a challenging goal, to monitor educational progress as educational priorities change. NAEP has reported unbroken trend lines for all 50 states, in Reading and Mathematics, Grades 4 and 8, for over 20 years. Unlike NAEP, state tests serve multiple purposes beyond measuring progress, including student feedback and school accountability. State tests are also influenced by shifting administrations of governors, commissioners, and legislatures.

In this symposium, we present new findings about State test trend lengths and magnitudes. The median state has had 3 different trendlines from 2009 through 2024. The typical state's trendline has an average length of 5 years, with considerable variability around this average. We introduce a new method for comparing these State trends to NAEP trends over the same periods, using random effects models to account for imprecision in both trendlines
Then, we bring national and state perspectives to these results. Suzanne Lane, a NAEP board member, will provide her perspective on the role of NAEP trends. Dan Farley, Oregon's director of assessment, will provide the state agency's perspective on the challenge of maintaining trend lines. Peggy Carr, who has long negotiated State-NAEP comparisons in her decades leading NCES, will serve as discussant.

Chair:
Benjamin Shear (University of Colorado Boulder)

Discussant:
Peggy Carr (Former Commissioner, National Center for Education Statistics, U.S. Department of Education)

Presentations:

1. **Broken Bridges: State Test Score Trend Lines Through Policies and Pandemics**
   Andrew Ho, Harvard University; Benjamin Shear, University of Colorado Boulder; Jie Min, Stanford University; Erin Fahle, Stanford University; Sean Reardon, Stanford University

2. **Precision-Adjusted Models for Comparing State and NAEP Test Score Trends**
   Sean Reardon, Stanford University; Andrew Ho, Harvard University; Benjamin Shear, University of Colorado Boulder; Erin Fahle, Stanford University; Jie Min, Stanford University

3. **The North Star? An Appropriate Role for NAEP in the Post-Pandemic Era**
   Scott Marion, Center for Assessment

4. **Challenges and Solutions in Maintaining and Explaining State Test Score Trends**
   Chris Rozunick, TEA

Data Monitoring: Empowering You to Be a Watch Dog and a Wizard

Coordinated Paper Session
11:30 AM – 12:45 PM
Intercontinental Los Angeles Downtown, Floor 5: Silver Lake A

Effective data monitoring is essential for maintaining the validity, reliability, and trustworthiness of educational assessment programs. A single oversight can undermine stakeholder confidence and damage an assessment program's reputation, sometimes requiring years to restore if recovery is possible at all. This coordinated session explores the critical role of data monitoring in educational measurement, emphasizing its dual capacity to serve as both a watchdog—catching issues before they escalate—and as a wizard—diagnosing and resolving problems after they occur. Presenters from Curriculum Associates, the Human Resources Research Organization (HumRRO), and WIDA share diverse perspectives and practical experiences in designing, implementing, and leveraging monitoring systems to proactively surface issues, support timely interventions, and strengthen the overall integrity of assessment programs.

The session opens with an overview of how data monitoring plans are developed at Curriculum Associates, followed by examples showing how approaches vary by product type, design, and stakes. WIDA presents a multi-layered data monitoring framework for ACCESS for ELLs, combining proactive quality control, forensic analytics, and responsive issue resolution. HumRRO provides examples of real-world assessment issues, solutions implemented, and recommends automated monitoring processes to catch issues. The session concludes with a summary and discussion of best practices for data monitoring.

Discussant:
Rich Patz

Presentations:

1. **Designing and Implementing Data Monitoring Plans—A Psychometric Blueprint**
   Kristin Morrison

2. **Monitoring In Motion: A Comparative Case Study Analysis in Assessment Monitoring**
   Nathan Minchen, Curriculum Associates; Sebastian Moncaleano-Wallrich

3. **WIDA: Safeguarding Assessment Integrity through Comprehensive Data Monitoring**
   Kyoungwon Bishop

4. **Assessment Data, Process Replication, and Quality Assurance: Before, During, and After**
   Arthur Thacker, Human Resources Research Organization

Deconstructing AI Outputs for Bias in the Workplace

Organized Discussion
11:30 AM – 12:45 PM
Intercontinental Los Angeles Downtown, Floor 5: Boyle Heights

Chair:
Montserrat Valdivia Medinaceli (Curriculum Associates)

Presenter(s):
Montserrat Valdivia Medinaceli, Curriculum Associates; Lissette Tolentino, University of Central Florida; Claudia Ventura, University of Connecticut; Catherina Villafuerte, University of Connecticut

Psychometricians are adopting AI language models for item ideation, coding assistance, literature triage, and reporting. Field practice privileges prompt craft over the interpretive reading of model outputs. This organized discussion centers on how to read AI responses with the same rigor we apply to data, models, and instruments. Using an explicitly justice oriented lens, we surface recurrent risks, including stereotyping, positional and verbosity effects, and language that resembles surveillance, and we connect these risks to consequences for fairness and institutional accountability in educational assessment. The format is fully participatory. We will poll current AI use, analyze work prompts submitted by the audience in real time, and collectively annotate model responses for bias patterns, followed by a brief synthesis of practical mitigations that participants can adopt in their teams. Attendees will leave with a compact checklist for interpreting AI outputs, examples from K to 12 and higher education assessment, and concrete strategies for embedding accountability into evaluation workflows.

Innovations in culturally and linguistically responsive measure development

Coordinated Paper Session
11:30 AM – 12:45 PM
Intercontinental Los Angeles Downtown, Floor 5 : Ladera Heights

There is a substantial history of racial, ethnic, cultural, and linguistic bias in assessment (Randall, 2021; Randall, et al., 2024; Solano-Flores, 2023). The practice of systematically deprioritizing historically marginalized children's identities and languages in assessment development and validation has yielded tools that inaccurately capture many children's abilities (i.e. Larry P. vs Riles). New approaches to measurement that strategically attend to the lived experiences and linguistic resources are needed. The Justice-oriented Antiracist Validation (JAV) framework can serve as a heuristic for guiding innovation to reduce bias and discriminatory assessment practices. (Randall, 2024). This symposium will include four presentations that challenge current assessment practices by exploring new approaches to gathering information from families to choose the best assessment match for young children, innovative approaches to evaluating the science knowledge of Spanish-English bilingual students, and analyzing reading screening prediction models that include both English and Spanish assessments to explore the best combinations of measures that maximize accuracy in predicting reading risk.

Discussant:
Michael Rodriguez (University of Minnesota)

Presentations:

1. **Applying a Justice-oriented Antiracist Approach to Early Childhood Assessment**
   Alisha Wackerle-Hollman, University of Minnesota; Lillian Duran, University of Oregon; Rebecca Nathan, Aviellah Curriculum Consulting

2. **Understanding Bilingual Students' Interpretation of Science Vocabulary Items: A Latent Class Analysis**
   Jose Palma, Texas A&M University; Noah Koehler, Texas A&M University; Holland Kowalkowski, University of Texas Austin

3. **Why Only Assess Half Their Skills? Predicting Multilingual's English Reading Risk from Their Skills in Both Spanish and English**
   Julian Siebert, University of California, San Francisco; Mónica Zegers, UCSF; Francesca Pei, University of California, San Francisco; Marilu Gorno Tempini, University of California, San Francisco

4. **Predicting Spanish reading risk in dual language programs using a cross-linguistic approach**
   Lillian Duran, University of Oregon; Julian Siebert, University of California, San Francisco; Francesca Pei, University of California, San Francisco; Marilu Gorno Tempini, University of California, San Francisco

Now what? Leveraging Use Cases for Assessment and Accountability Design and Data

Organized Discussion
11:30 AM – 12:45 PM
Intercontinental Los Angeles Downtown, Floor 5 : Silver Lake B

Chair:
Aneesha Badrinarayan (Education First)

Discussant:
Susan Lyons (Lyons Assessment Consulting)

Presenter(s):
Aneesha Badrinarayan, Education First; Nathan Dadey, The National Center for the Improvement of Educational Assessment; Cedar Rose, Montana Office of Public Instruction; Kyu-Ryng Hwang, New Hampshire Department of Education

Many assessment programs are created around broad goals such as "program evaluation" or "informing instruction." While well-intentioned, this high-level framing often leaves educators and leaders with assessments that are difficult to connect results to concrete actions at the school, district, or state level. A more effective approach may be to begin with clearly defined, specific intended and likely uses of an assessment program, design for those specific uses, and then revisit those uses throughout the program's lifecycle–from design through implementation. Misalignment between intended and actual uses helps explain why some assessment programs fail to generate productive responses. Yet this work is not simple: defining uses at the right level of detail, balancing the diverse priorities of multiple stakeholders, and continuously updating the program all introduce challenges and trade-offs that need to be navigated carefully. This organized discussion focuses on how innovative assessment systems are leveraging principles of human-centered design alongside rigorous assessment and accountability system elements to transform how large-scale measurement systems are designed and implemented for impact. This session examines concrete examples and practical strategies for design and implementation of large-scale assessment systems.

The Design Dialog: How AI Amplifies and Needs Assessment Design Frameworks

Organized Discussion
11:30 AM – 12:45 PM
Intercontinental Los Angeles Downtown, Floor 7: Roosevelt B
Chair: Kristen Huff (Curriculum Associates)

Discussant:
John Behrens (University of Notre Dame)

Presenter(s):
Kristen Huff, Curriculum Associates; Kristen Dicerbo, Khan Academy; Britte Cheng, Menlo Education Research; John Behrens, University of Notre Dame

This panel explores how AI has the potential to revolutionize the application of evidence-centered design (ECD) and principled assessment design (PAD), through automated generation of design artifacts and test-taking elements. We also explore how ECD & PAD are more important for AI-augmented development than in traditional scenarios, as they provide a design language with the specificity needed to appropriately guide generative AI system behavior. These possibilities also raise new questions about validity, transparency, and equity that will be discussed.

- Kristen Huff, Curriculum Associates will describe how large language models (LLMs), when trained on high quality items, curated metadata, and PAD artifacts can potentially bypass the need for task model development without compromising construct clarity or item quality.
- Kristen DiCerbo, Khan Academy will examine how AI can augment each stage of the ECD framework—student model, evidence model, and task model—expanding both the scale and flexibility of assessment design.
- Britte Cheng, Menlo Education Research will situate AI within the broader ecosystem of learning sciences and equity, emphasizing the importance of human-centered design choices.

Moderator John Behrens, University of Notre Dame will engage the panelists and audience in dialogue around the opportunities and cautions for use of AI in assessment implementation.

Validity and Consequential Evidence: Moving Measurement Forward for Alternate Assessments

Coordinated Paper Session
11:30 AM – 12:45 PM
Intercontinental Los Angeles Downtown, Floor 6: Majestic

This session explores advancements in assessment measurement, with a specific focus on alternate content and English language proficiency assessments. This set of papers describes research and advances being made in four alternate assessment systems (two academic, two English language proficiency). The assessment systems are at various stages of maturity. The moderator will set the stage to address important measurement topics and advancements for both alternate content assessments and alternate English language proficiency assessments situated within a set of legislative policies and regulations.

The papers demonstrate innovative, equitable, and inclusive practices that advance measurement science and better inform teaching, learning, and decision-making for students who take these tests. A key focus of this session is the critical, yet often-overlooked, role of consequential validity. This aspect of validity examines the influence of test scores on important decisions and outcomes for students. By using an evidentiary process for continuous improvements, papers from two alternate assessment consortia and two English language proficiency consortia illustrate how they are addressing issues related to meeting test purpose and use. Through this lens, the session provides crucial insights into how measurement can drive positive, intended consequences, and address unintended negative outcomes for vulnerable student populations.

Chair:
Melissa Gholson (ATLAS, University of Kansas)

Discussant:
Anne Davidson (CrescendoEd LLC)

Presentations:

1. **Challenges of a Balanced Assessment System: Alternate Interim Assessments for Students with Significant Cognitive Disabilities**
   Bethany Spangenberg, Deputy Associate Superintendent of Assessment

2. **Challenges and Innovations in Developing and Validating an ELP Assessment for an Alternate Population**
   Eunhee Keum, ELPA21 at UCLA CRESST; Jenny Kao, ELPA21 at UCLA CRESST

3. **Building Validity Evidence for WIDA's Alternate ACCESS: From Foundational Research to Alignment**
   Laurene Christensen, WIDA at the University of Wisconsin-Madison

4. **Alternate Assessment Participation and Performance over the Years: ESSA-era Changes in State Policy and Practice**
   Meagan Karvonen, ATLAS, University of Kansas ; Melissa Gholson, ATLAS, University of Kansas; Nami Shin, ATLAS, University of Kansas

NCMEntoring: Fireside Chat (Invited Session: Membership Committee)

Organized Discussion
11:30 AM – 1:00 PM
Intercontinental Los Angeles Downtown, Floor 5 : K-Town

Are you a graduate student or an early career professional? Are you interested in an opportunity to meet and mingle with more experienced researchers and practitioners in the field? If so, please join us for a fireside chat as part of the NCMEntoring program. You will hear from a selected group of "mentors" with diverse experience in our field during a panel discussion in the first part of the session and then be able to informally network with them during table group discussions in the second part of the session. This session is open to all conference participants whether or not you have signed up for the 1:1 mentoring opportunities through the NCMEntoring program.

Chair:
Laurie Davis (Curriculum Associates)

Fireside Chat Speakers:
Dena Pastor, James Madison University; Susan Lottridge, Pearson; Matthew Madison, University of Georgia; Andre Rupp, Center for Assessment; Katherine Furgol Castellano, ETS Research Institute

Using AI for Aligning Standards

Coordinated Paper Session
11:30 AM – 1:00 PM
Intercontinental Los Angeles Downtown, Floor 7: Hollywood Ballroom II

Large Language Models (LLMs) are poised to revolutionize the labor-intensive processes of item development and standard alignment in large-scale assessment. Alignment tasks have relied on manual, time-consuming efforts by subject matter experts (SMEs), often leading to inconsistencies. LLMs solve this by leveraging their pretrained understanding of text. They can identify subtle semantic relationships between assessment items and educational standards, or between different sets of standards, dramatically improving both the efficiency and consistency of the alignment process. This research promises more valid, reliable, and efficiently developed assessments.
This session features researchers from Cambium Assessment, Centriverse, WestEd, HumRRO, and the University of Massachusetts Amherst showcasing a range of LLM-based approaches to standard-setting and alignment. These approaches include the use of embeddings, LLMs fine-tuned for classification, and prompting and fine-tuning strategies for generative LLMs, providing a comprehensive, multi-disciplinary approach to item development and standard setting.

Discussant:
Karla Egan (EdMetric)

Presentations:

1. **Trials, Tribulations, and Successes in Using AI to Align Adult Literacy Standards and Workplace Skills**
   Fernando Mena, University of Massachusetts Amherst; Stephen Sireci, University of Massachusetts Amherst; Javier Suárez-Álvarez, University of Massachusetts, Amherst; Anna Sullivan, University of Massachusetts Amherst

2. **Expanding Validity Evidence with Large Language Models: Alignment, Standards Mapping, and Test Item Transferability**
   Harold Doran, HumRRO

3. **Enhancing LLM-Assisted Alignment with Semantic Augmentation**
   Hye-Jeong Choi, HumRRO; Reese Butterfuss, Certiverse; Michael Kelly

4. **Fine-tuning of Large Language Models to Align Items to Standards**
   Suhwa Han, Cambium Asessment; Christopher Ormerod; Frank Rijmen, Cambium Assessment

5. **AI-Supported Standard Setting Using SMART Methodology with NAEP Items**
   Sarah Quesen, WestEd

Classroom Assessment Committee
Meeting
12:30 PM – 1:30 PM
Intercontinental Los Angeles Downtown, Floor 6: Royal

The purposes of the Classroom Assessment Committee are to enhance the equitable and high-quality practice of classroom assessment, to enhance the visibility and awareness of classroom assessment among the NCME membership, and to spur and communicate classroom assessment scholarship beyond the NCME membership by promoting research and practice.

Innovation Demonstrations
1:45 PM – 2:45 PM
Intercontinental Los Angeles Downtown, Floor 5 : Hancock Park

Presentations:

1. **A Procedure for Item Compromise and Preknowledge Detection in Computerized Adaptive Testing**
   Yichong Cao, Delaware Department of Education; Jianshen (Cassie) Chen, College Board; Luz Bay, Independent Researcher

   A R-based analytic system is developed to detect compromised items and examinees with preknowledge using item score and item response time in computerized adaptive testing. The method, analytic system and results of the proposed three-step procedure will be presented, and code will be shared upon request.

2. **Guiding field testing with text-based item difficulty prediction using ML classification models**
   Xiuhan Chen, University of Missouri; Xuechun Zhou, Ascend Learning

   The time-consuming nature inherent in field testing presents significant obstacles to supplementing an item bank as needed timely. Aiming to develop a guided approach to field testing, this study investigates the accuracy of text-based item difficulty prediction in the domains of Math and Science using machine learning classification models.

3. **mstATA: An R Package for Automated Multistage Test Assembly**
   Hong Chen, University of Iowa; Anthony Fina, University of Iowa; Jing Ma, University of Iowa; Catherine Welch, University of Iowa; Stephen Dunbar, University of Iowa

   This paper introduces an R package for automated test assembly using mixed-integer linear programming (MILP). It is specifically designed for multistage testing and supports top-down, bottom-up, and hybrid strategies. This package provides transparency and flexibility in test construction while leveraging modern solver technologies to address large-scale test assembly challenges.

4. **Estimating and Evaluating Diagnostic Models with the R Package measr**
   Jake Thompson, ATLAS, University of Kansas

   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.

5. **Multi-Method Cheating Detection with Voting Strategy**
Yunting Liu, University of California, Berkeley; Irina Grabovsky; Richard Feinberg, NBME; Chunyan Liu, National Board of Medical Examiners

We compare three kinds of cheating detection methods, K-index, mixture model, and multiple person fit indices. Their pros and cons are discussed with evidence from simulation studies. Acknowledging the divergence in classification results, we propose a voting strategy that improves the robustness of detection power while controlling Type I error.

6. **Scenario-Based Assessment for AI Literacy in High School Students**
Caitlin Tenison, ETS; Michael Suhan, Educational Testing Service; Tenaha O'Reilly

In this demonstration, we introduce a scenario-based assessment prototype for AI literacy. Designed for high school students, the task uses a chatbot interface to measure responsible and critical AI use. We highlight design challenges and solutions for balancing standardization with authentic interaction and invite attendees to explore key prototype features.

7. **An LLM-Augmented Score Reporting Assistant for K-12 Teachers**
Shan Zhang, University of Florida; Caitlin Tenison, ETS; Diego Zapata-Rivera, ETS; Reginald Gooch, ETS

This project develops an LLM-powered Smart Report Assistant, grounded in audience analysis and assessment design principles, that uses retrieval and visualization to help K-12 teachers interpret assessment data at classroom, group, and individual levels. The system highlights key learner dimensions and supports data-driven pedagogical strategies through personalized and conversational reporting.

8. **Advancing Multistage Test Evaluation with the Expected Test Information Function**
Cheng Hua, University of Montevallo; Liuhan Cai, Accenture; Louis Roussos, Cognia

This demonstration introduces upgraded software for applying the Expected Test Information Function (ETIF) in multistage test evaluation. Enhanced features include data import/export, multi-path comparisons, and advanced visualization tools. Using real-world application scenarios, we demonstrate how ETIF enhances psychometric practice by simplifying analysis, strengthening communication, and improving test development efficiency.

9. **Analyzing Student Cohorts with cohortED in R: A Reproducible, Modular Approach**
Brian Harrold, University of South Carolina; Nathan Dadey, The National Center for the Improvement of Educational Assessment; Damian Betebenner, Center for Assessment

This demonstration introduces cohortED, an R package for analyzing student cohorts across schools, districts, and states. The package produces R objects for modular analysis and generates dynamic Quarto reports, enabling analysts to clean, analyze, and communicate longitudinal patterns in composition, performance, and mobility as reproducible insights and ready-to-use reports.

10. **Automated Test Assembly Using R Packages for Tests with Testlets**
Huijuan Wang; Keith Boughton, Data Recognition Corporation; Jessalyn Smith, DRC

This study introduces two R packages, ROI and ompr, to implement Automated Test Assembly (ATA) for tests with testlets. Both packages efficiently handle ATA tasks, with ompr producing more balanced test forms. The study highlights the practical implications for researchers and practitioners, offering accessible tools for assembling complex, high-quality assessments.

11. **PrereX - Conducting Expert Queries for Hierarchical Competency Structures Automatically.**
Peter Steiner, St.Gallen University of Teacher Education; Jan Hochweber, St. Gallen University of Teacher Education, Switzerland; Yan Wang, St.Gallen University of Teacher Education; Fabian Grünig, St.Gallen University of Teacher Education; Stephanie Leininger, St. Gallen University of Teacher Education, Switzerland; Michael Kickmeier-Rust, St.Gallen University of Teacher Education; Stephan Schönenberger, St. Gallen University of Teacher Education, Switzerland

This demonstration introduces an open-source web application for constructing competency structures based on expert queries. The system automatically generates and optimizes queries from a competency domain, supports discrepancy detection, and exports precedence matrices, thereby facilitating the practical application of psychometric models grounded in hierarchical structures such as KST or AHM.

12. **Digital Test Forms Selection Tool for Item Matrix Sampling Design: NAEP Application**
Mathew Kandathil, ETS Research Institute; Edward Kulick, ETS Research Institute; Paul Hilliard

Item matrix sampling is a test design approach to control testing time while maintaining board content coverage. The approach involves dividing a set of items into multiple forms. We will illustrate an innovative approach to form assignments for digital tests, which greatly improved the precision and efficiency of form assignments.

13. **Introducing an LLM-Powered Chatbot for Supporting Accommodation Development for Learning & Assessment**
Magdalen Beiting-Parrish, EdAIfy, CUNY Graduate Center

Approximately 15% of public-school students are Students with Disabilities and 10% are Multilingual Learners. Both tend to underperform relative to their peers and need accommodations to perform best. This demonstration will present an LLM-based chatbot that will assist educators and test developers in creating accommodations for classroom and assessment purposes.

14. **py-equate: A Python Package for Flexible Score Equating**
Laura Lambert, James Madison University; Yu Bao, James Madison University

This demonstration introduces py-equate, a Python package that integrates observed-score and IRT-based equating methods into one accessible toolkit. Designed for psychometricians and doctoral-level instruction, py-equate offers user-friendly syntax and visualizations. Attendees will learn how to install the package, apply methods, and interpret equating results.

15. **AI-Driven Certification Exam Item Generation Using Multi-Agent Workflows**
Vinita Talreja

We will demonstrate an automated system that transforms skill statements from an exam outline into high-quality certification exam questions using coordinated AI agents. Combining automated research, content synthesis, and quality evaluation, it generates scenario-based multiple-choice items with verified technical accuracy, offering a scalable solution for item development across technical domains.

16. **Conversational AI for Education Data: Beyond Dashboards to Natural Language Inquiry**
Erik Whitfield

This demonstration introduces a terminal-based conversational AI system that replaces traditional education data dashboards with natural language queries. Using Retrieval-Augmented Generation (RAG), users can explore school performance data through intuitive questions rather than navigating complex interfaces, making education analytics accessible to all stakeholders.

17. **PsychMet: An AI-Powered Chatbot for Measurement Competencies**
Henry Makinde; Mubarak Mojoyinola; Hope Adegoke, University of North Carolina, Greensboro

PsychMet is an AI-powered chatbot built on GPT-4.1 with Retrieval-Augmented Generation (RAG). It supports personalized development of NCME foundational competencies by delivering contextualized explanations, curated resources, and adaptive feedback. The demonstration highlights PsychMet's design, evaluation using RAGAS, and its potential to transform professional development in educational measurement.

AI Ethics, Policy, and Practice in Training Measurement Professionals: A Panel Discussion (Educators of Measurement SIGIMIE Session)

Organized Discussion
1:45 PM – 3:15 PM
Intercontinental Los Angeles Downtown, Floor 7: Roosevelt A

Chair:
Justin Kern (University of Illinois at Urbana-Champaign)

Discussant:
Victoria Quirk (University of Illinois at Urbana-Champaign)

Presenter(s):
Justin Kern, University of Illinois at Urbana-Champaign; Hong Jiao, University of Maryland; Okan Bulut, University of Alberta; Christopher Runyon, National Board of Medical Examiners; Joe Grochowalski, College Board; Michael Rodriguez, University of Minnesota

The rapid rise of generative artificial intelligence (genAI) has created both opportunities and challenges across academic and professional domains. Within educational measurement, conversations have largely centered on how genAI can support professional work, such as test item generation, automated coding, or streamlining research processes. However, less attention has been given to how measurement professionals themselves should be trained to use, evaluate, and critique these tools, particularly with respect to ethics, policy, and practice. This panel, organized by the Educators of Measurement (EoM) SIGIMIE, will bring together experts from academia and industry to discuss how the educational measurement community should address genAI in the preparation of future professionals in both academia and research and testing organizations. Panelists will explore ethical considerations, training needs, and the evolving competencies required for measurement practitioners, with a special focus on alignment with NCME's Foundational Competencies in Educational Measurement.

Assessing Invariance of Automated Scoring Models

Coordinated Paper Session
1:45 PM – 3:15 PM
Intercontinental Los Angeles Downtown, Floor 5 : K-Town

Automated scoring models are increasingly used in high-stakes educational assessments, necessitating rigorous evaluation of their reliability, fairness, and generalizability across diverse contexts. This coordinated paper session presents five studies that address critical aspects of invariance in automated scoring. The first paper introduces an entropy-aware generalizability theory framework to quantify and optimize reliability of transformer-based scoring models under cost constraints. The second paper investigates the impact of training data composition on predictive accuracy and fairness, focusing on subgroup representation in transformer-based models. The third paper extends empirical Bayesian fairness assessment by jointly modeling group-level mean differences and implementing multivariate loss functions for improved item-level decisions. The fourth paper evaluates the operational generalizability and fairness of hybrid scoring models across states and subjects in a large-scale assessment program. The fifth paper explores individual fairness in automated scoring by operationalizing rubric-relevant feature similarity, aiming to ensure equitable scoring for diverse responses. Collectively, these papers advance the understanding of reliability and fairness in automated scoring, providing practical methodologies for monitoring and improving model performance in educational measurement.

Chair:
Edward Wolfe (Iowa Testing Programs / University of Iowa)

Discussant:
Susan Lottridge (Pearson)

Presentations:

1. **Evaluating the Generalizability of Automated Scoring Models to a State-Level Assessment Context**
   Corey Palermo, Measurement Incorporated

2. **From Entropy to Generalizability: Strengthening Automated Essay Scoring Reliability and Sustainability**
   Yi Gui, Measurement Incorporated (MI)

3. **Bias and Accuracy Evaluations of Large Language Models in Automated Essay Scoring**
   YoungKoung Kim, College Board; Tim Moses, Buros Center for Testing; Daniel Lee, College Board

4. **An Approach for Assessing Individual-Level Fairness in Automated Scoring**
   Matthew Johnson, ETS Research Institute; Michael Fauss, ETS; Ikkyu Choi, ETS

5. **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, BroadMetrics

Assessment Needs: Can We Build Balanced Systems of Assessment to Address Them? (Classroom Assessment Committee)

Organized Discussion
1:45 PM – 3:15 PM
Intercontinental Los Angeles Downtown, Floor 7: Roosevelt B

Chair:
Duy Pham (New Meridian Corporation)

Discussant:
Duy Pham (New Meridian Corporation)

Presenter(s):
Duy Pham, New Meridian; Stephen Sireci, University of Massachusetts Amherst; Joseph Sireci, Capitol Hill High School, Oklahoma City, OK; Julie Goldberg, Poway Unified School District, San Diego, CA; Valentyna Banner, San Diego Global Vision Academy, San Diego, CA

Calls for balanced systems of assessment that support both instruction and student learning—have persisted for decades. Yet, implementation remains uneven across K–12 education. This organized discussion brings together a teacher, a school administrator, a district leader, and a psychometrician to explore how assessment systems can better meet the needs of students, educators, and administrators. The session will unfold in three parts. First, panelists will share their perspectives on the most pressing assessment needs at their respective levels and discuss two key barriers to building balanced systems. Second, they will evaluate how current assessments address these needs, highlighting successes, gaps, and persistent challenges. Finally, each panelist will propose actionable recommendations to overcome the identified barriers and strengthen assessment systems. The discussion will focus on two major barriers: (1) the influence of politics and policy, and (2) the lack of attention to curriculum and learning in assessment design and development. Audience participation will be encouraged, and key takeaways will be summarized in a post-conference blog to extend the conversation beyond the session.

Estimation and Scoring

Individual Paper Session
1:45 PM – 3:15 PM
Intercontinental Los Angeles Downtown, Floor 5: Silver Lake A

Chair:
Bob Schwartz (National Conference of Bar Examiners)

Discussant:
Tony Albano (University of California - Davis)

Presentations:

1. **Modified Approaches to Classification Consistency and Accuracy for Weighted IRT Scoring**
   Audrey Filonczuk; Seohee Park, American Board of Internal Medicine

   In longitudinal assessments, recent item responses may be weighted more heavily to reflect examinee growth, introducing scoring methodology not yet examined in classification consistency and accuracy literature. We adapt three IRT-based approaches, Rudner (2001, 2005), Guo (2006), and Lee (2002, 2010), to incorporate weighted scoring schemes and validate them empirically.

2. **Summed Scores for Prediction: Model-Implied and Conformal Intervals**
   Youmin Hong, University of Maryland; Youngjin Han, University of Maryland; Yang Liu, University of Maryland; Youjin Sung, University of Maryland; Ji Seung Yang, University of Maryland

   For constructs measured by multiple indicators, a debate exists: should summed-score regression be replaced by latent variable models or machine learning techniques? Our simulations show that, when paired with conformal prediction intervals, summed-score regression remains a robust tool that can outperform more complex predictive models when they are misspecified.

3. **Introducing the GMUPP IRT Model for Graded Multidimensional Forced Choice Measures**
   Lavanya Shravan Kumar, Graduate Management Admission Council; Sean Joo, University of Kansas; Stephen Stark, University of South Florida

   Graded multidimensional forced choice (MFC) measures are emerging as a promising alternative to binary MFC measures, rating scales in noncognitive assessment. This research develops the GMUPP, a generalization of the Multi-Unidimensional Pairwise Preference (MUPP) model, for graded MFC responses, and compares latent trait estimation assuming three models for single-stimulus responding.

4. **Global Maximum Is Not Always Best in 3PL Maximum-Likelihood Scoring: Smarter Initialization**
   Hwanggyu Lim, Inha University; Edison Choe, Renaissance Learning; Kyung (Chris) Han, GMAC

   This study examines whether selecting the global maximum is always optimal for 3PLM maximum-likelihood scoring. Simulations show that interior maxima can yield greater accuracy than terminal solutions at boundaries. We also introduce Smart-Grid-Search (SGS), an efficient initialization strategy that targets promising interior maxima to improve scoring accuracy and efficiency.

5. **Concurrent vs Fixed Item Parameter Calibration: Impact on Equated Cut Scores**
   Derya Cakici-Eser, Pearson; Anna Sullivan, University of Massachusetts Amherst; Amy Schmidt, Pearson

   This study examines the impact of concurrent and fixed item parameter calibration methods on equated cut scores within an IRT framework. Using operational data from a teacher licensure exam, the analysis reveals significant differences in item parameter estimates and equated cut scores, offering practical insights for psychometricians.

Frameworks and Considerations for Dimensionality Assessment

Individual Paper Session
1:45 PM – 3:15 PM
Intercontinental Los Angeles Downtown, Floor 6: Majestic

Chair:
Zachary Mayne (IXL)

Discussant:
Yoojoong Kim (University of Georgia)

Presentations:

1. **Corroborative Factor Analysis**
   Wes Bonifay, University of Missouri; Sonja Winter, University of Missouri

   To bring measurement modeling into closer alignment with important philosophical-scientific principles (e.g., severe testing, the weight of evidence, and theory corroboration), we introduce corroborative factor analysis: an inferential framework for (dis)corroborating factor analytic theories via constrained maximum likelihood estimation and/or Bayesian prior predictive model checking.

2. **Refactoring Analyses: Direct Tests of Dimensionality from Indirect Data**
   Michael Hardy, Stanford University

   Traditional unidimensionality tests are indirect, evaluating fit on statistical images rather than the data itself. Refactor Analyses evaluate low-rank model reconstructions of the observed data. Experiments across 50 public datasets reveal that traditional metrics weakly correlate with unidimensional model capacity to capture original response variation, questioning traditional assumptions.

3. **Using Classification Trees for Method Selection from Observed Features in Simulation Studies**
   Jeongwon Choi, Vanderbilt University; Hao Wu, Vanderbilt University

   Monte Carlo simulations compare methods using fixed population values, while researchers only have access to observable data features. We propose a framework that samples across continuous parameters and translates results into decision rules with classification trees, illustrated through zero-cell corrections for tetrachoric correlations and consideration of alternative data representations.

4. **Handling Missing Data in EFA: Parametric and Nonparametric Bootstrapping with Multiple Imputation**
   Xinchang Zhou; Ruoqian Wu, University of Illinois Urbana-Champaign; Yan Xia, University of Illinois at Urbana-Champaign

   This study uses a Monte Carlo simulation to systematically compare the MI-Bootstrap and Bootstrap-MI workflows for assessing EFA stability with missing data. Results indicate that MI-Boot, where imputation precedes bootstrapping, provides superior accuracy and stability. Findings provide researchers an evidence-based guideline for obtaining more reliable EFA results from incomplete data.

5. **A Systematic Evaluation of the Effects of Careless Responding on Exploratory Factor Analysis**
   Tugay Kaçak, Trakya University; Abdullah Kılıç, Trakya University

   This study uses Monte Carlo simulations to examine how different types of careless responding distort exploratory factor analysis. Findings show biases in factor loadings and inter-factor correlations vary by careless respondent type and magnitude of factor loadings, offering critical guidance for scale development and psychometric research.

From Guessing to Carelessness: Examinee Effort and Its Impact in Assessment Contexts

Coordinated Paper Session
1:45 PM – 3:15 PM
Intercontinental Los Angeles Downtown, Floor 5: Westwood

Expended examinee response effort is a critical yet often overlooked factor influencing the validity of inferences from assessment scores. This session explores how disengaged responding can be measured and decreased across a variety of different assessment environments and respondent populations. Through diverse methodologies and low-stakes data collection contexts, the papers in this session advance assumptions about disengagement, revealing that very slow responses may reflect thoughtful engagement and that multi-indicator approaches uncover nuanced profiles of disengaged behavior. Additionally, findings show that rapid responding on risk assessments can misclassify high school students' risk levels, with implications for educational decision-making. Strategies to improve effort, such as refining test administration procedures and strategically shortening assessments, demonstrate potential to enhance both engagement and performance. Together, these studies underscore the importance of detecting and addressing non-effortful responding to ensure fair, reliable, and valid use of assessment data. Attendees will gain practical insights into identifying disengagement and implementing design and procedural improvements that promote meaningful participation in low-stakes testing environments.

Discussant:
James Soland (University of Virginia)

Presentations:

1. **Validating Slow Responses as Indicators of Disengaged Test Taking: Not So Fast**
   Steven Wise, EngagedMeasurement; Dena Pastor, James Madison University; Tzu-Chun Kuo, Kaplan North America; Nina Deng

2. **Identifying Classes of Careless Responders Using Multiple Methods**
   Katarina Schaefer, James Madison University; Deborah Bandalos, James Madison University

3. **Disengaged Responses, Distorted Risks: Evaluating Effort in Student Risk Assessments**
   Laura Pires Gifford, Washington State University; Riley Herr, James Madison University; Dakota Galangue, University of Delaware; Brian French, Washington State University; Sara Finney, James Madison University

4. **Assessing Examinee Effort: The Role of Test Length in Low-stakes Testing**
   Jonathan Henriques, James Madison University; Brian Leventhal, James Madison University

IRT Analyses with Person and Item Covariates

Individual Paper Session
1:45 PM – 3:15 PM
Intercontinental Los Angeles Downtown, Floor 5 : Boyle Heights

Session Description:
Chair: Kuan-Yu Jin (Hong Kong Examinations and Assessment Authority)

Discussant:
Michael Peabody (IXL Learning)

Presentations:

1. **Explanatory Item Response Theory for Specialty-Domain Alignment in Medical Recertification Exam**
   He Ren, University of Washington; Yanlin Jiang, NCCPA; Andrew Dallas, NCCPA; Mirela Bruza, NCCPA

   This study employs a cross-classified explanatory item response theory (EIRT) model to investigate how alignment between test-taker's clinical specialties and the content domains of exam items influences performance in medical assessments. The analysis quantifies the specialty-domain alignment effect and evaluates its implications for exam scoring practices and validity evidence.

2. **The Construct Validity of Multiple-Choice Items in the HKDSE Mathematics Exams**
   ZHEN WANG, Hong Kong Examinations and Assessment Authority; Kuan-Yu Jin, Hong Kong Examinations and Assessment Authority; Joseph Chow, Hong Kong Examinations and Assessment Authority

   This study examined the construct validity of the multiple-choice items in the HKDSE mathematics tests with the Rasch model and the linear logistic test model. It indicated that item difficulties can be appropriately explained by six content domains and five cognitive domains. The results were validated via parallel analysis.

3. **Distillation of Mixture IRT Model with Response Time to Classify Examinees' Behaviors**
   Bowen Wang, National Board of Chiropractic Examiners; Igor Himelfarb; Nai-En Tang, National Board of Chiropractic Examiners

   This study extends mixture IRT models to classify heterogeneous examinee behaviors in a high-stakes licensure exam. A three-latent-class mixture model with response time identifies efficient, careful, and inefficient groups. Using model distillation, random forest replicates classifications with high accuracy, offering generalizable classification and real-time behavioral monitoring in large-scale assessments.

4. **Capturing Responses Process Variation with A Within-person Mixture IRTree Model**
   Jinwen Luo, UCLA; Sijia Huang, Indiana University Bloomington; Minjeong Jeon, UCLA

   While completing self-reported surveys, respondents may show different response styles, and may not follow the same response process across items. To account for both between- and within-person variation in response processes, we introduce a within-person mixture IRTree model. We evaluate the model via simulation and empirical data analysis.

5. **Disengagement in Testing: An IRTree Model for Change Over Time**
   Brian Leventhal, James Madison University

   Test scores reflect both ability and test-taking effort, especially in low-stakes assessments. In this study, I present a longitudinal extension of the IRTree model for Disengagement that adjusts ability and disengagement over time, improving the accuracy of growth estimates from educational assessments.

Methodological Issues in Growth Modeling

Individual Paper Session
1:45 PM – 3:15 PM
Intercontinental Los Angeles Downtown, Floor 5 : Ladera Heights

Chair:
Ze Wang (Amazon Web Services)

Discussant:
Ze Wang (Amazon Web Services)

Presentations:

1. **A Comparison of Machine Learning Approaches to Longitudinal Data**
   Hyewon Chung, Chungnam National University; Eunah Jang

   This simulation study compared mixed-effects random forest (MERF) and generalized linear mixed models with Lasso regularization (glmmLasso), both introduced to account for longitudinal data structures. Results showed glmmLasso consistently outperformed MERF. MERF's RMSE increased with more predictors, more repeated measurements, and smaller samples, whereas glmmLasso remained relatively stable across conditions.

2. **Comparing Methods for Estimating Conditional Growth Percentiles for an Interim Assessment**
   Luciana Cancado, Curriculum Associates; Logan Rome, Curriculum Associates; Jiawei Xiong, Curriculum Associates

   This study compares three methods for estimating conditional growth percentiles with interim assessment data: more established Student Growth Percentiles and Percentile Rank of Residuals, and a related but less explored method we call Theoretical Gain Percentiles. Results show that percentile ranks assigned by all methods are highly correlated and consistent.

3. **Using Mixed-effects Random Forest to Analyze Large-scale Assessment Data with Hierarchy**
   Miryeong Koo, University of Illinois at Urbana-Champaign; Jinming Zhang, University of Illinois at Urbana-Champaign

   This study aims to propose methods to automatically explore optimal hierarchical linear models, with machine learning techniques, from any large-scale assessment data with hierarchy. The result shows the method with mixed-effects random forest outperforms the one with random forest and its superiority becomes more prominent with higher intraclass correlation coefficients.

4. **Modeling Growth in Progress Testing: A Multilevel Analysis of Student and School effects in CBSE Performance**
   Irina Grabovsky; Jerusha Henderek, NBME; Francis O'Donnell, National Board of Medical Examiners

   We model CBSE progress testing with three-level HLM (occasions within students within schools), using Weeks and Attempt0. Scores rise as Step 1 nears; additional attempts raise levels and steepen time trends. Likelihood-ratio tests favor the model. Variance decomposes across school, student, occasion. Findings inform attempt policy and longitudinal measurement practice.

5. **The longitudinal trajectory of absenteeism for Minnesotan high school students across 9 cohorts**
   Shandell Pahlen

   We investigated absenteeism growth curves for Minnesotan public high school student across 9 cohorts. Overall, absenteeism tends to gradually grow across time, peaking by senior year. Differential cohort patterns persist highlighting the impact of state accountability policies and covid. Student and school factors on absenteeism trajectories will be further examined.

Modeling Approaches for DCM

Individual Paper Session
1:45 PM – 3:15 PM
Intercontinental Los Angeles Downtown, Floor 5 : Silver Lake B

Chair:
Christoper Ocheni (The University of Alabama, Tuscaloosa)

Discussant:
Wenchao Ma (University of Minnesota)

Presentations:

1. **Evaluating 1PL-CDM Test Blueprints Robustness to DIF**
   Nicolas Mireles; Matthew Madison, University of Georgia

   In this study, we evaluate whether differential item functioning distorts empirical blueprints in the 1PL-CDM. We manipulated DIF type, magnitude of DIF, prevalence of DIF, group mastery impact, and group balance across 108 conditions. Blueprint bias was near zero, even at extreme conditions. Results support robustness of blueprints from the

2. **Nonparametric CD–CAT for Polytomous Attributes**
   Youn Seon Lim, University of Cincinnati

   NP–CAT–Poly is a nonparametric adaptive algorithm for cognitive diagnosis with polyto-mous attributes and multiple-choice options. It avoids calibration, integrates a Q-optimal start, and classifies via distance-based discrimination. Simulations and a classroom dataset show higher accuracy and shorter tests than baselines, supporting practical small-sample deployment.

3. **A mixture CDM for mathematical problem-solving strategies**
   Teng WANG; Xiangdong Yang, East China Normal University; Jing Huang, Purdue University

   This study proposes a mixture cognitive diagnosis model (M-CDM) for math problem-solving strategies from the perspective of cognitive processes, which analyzes and models solution behavior within students' problem-solving strategies. The findings indicate that the model can effectively identify and diagnose problem-solving strategies among accuracy rates and response time.

4. **Identifying Salient Attribute- and Item-level Covariates in Bayesian DCMs with Attribute Hierarchies**
   Alfonso Martinez, Fordham University; Fabio Setti, Fordham University; Zhixing Liu, Fordham University

   In this study, we utilize Monte Carlo simulation studies to study Bayesian explanatory diagnostic classification models with variable selection priors (e.g., Bayesian Lasso, continuous/discrete spike-and-slab priors) for identifying salient predictors of attribute mastery status and response behaviors in the presence of attribute hierarchies, including linear, convergent, and divergent attribute structures.

5. **On the Robustness of the One-parameter Log-linear Cognitive Diagnosis Model**
   Matthew Madison, University of Georgia; Oluwatosin Adeleye, University of Georgia; Nancy Alila, Department of Educational Psychology, University of Georgia, Athens, GA, USA; Ashley Heady; Beltran Pantoja, University of Georgia

   The one-parameter log-linear cognitive diagnosis model (1-PLCDM) has attractive measurement properties and the ability to match pre-specified test blueprints, which is crucial for operational application. The 1-PLCDM, however, makes strong assumptions to achieve these properties. This study examines the performance of the 1-PLCDM when its assumptions are violated.

Certification & Licensure SIGIMIE

Meeting
1:45 PM – 4:00 PM
Intercontinental Los Angeles Downtown, Floor 6: Royal

Certification and licensure is an important application of educational measurement used to protect the public by holding professionals to minimum standards within their respective fields.
The assessments used in certification and licensure have some of the highest stakes in the measurement community, and both leverage and are influenced by NCME's output. The purpose of this NCME SIGIMIE is to a) create space at the NCME conference for certification and licensure organizations to share their current work and identify common concerns, and b) provide a clear connection point for experts throughout NCME membership to understand current research opportunities in the space.

Beverage Break

Social
3:00 PM – 4:00 PM
Intercontinental Los Angeles Downtown, Floor 5 : Wilshire Grand Prefunction

Hot coffee & tea, lemonade, and iced tea will be available in the Wilshire Grand Prefunction area.

Graduate Student eBoards: Process data, CAT, technology, and large-scale assessment

Graduate Student Electronic Board Session
3:45 PM – 4:45 PM
Intercontinental Los Angeles Downtown, Floor 5 : Hancock Park

Presentations:

- **Time-Series Clustering of Response Times in Low-Stakes Computerized Adaptive Testing**
  Minjung Kim, Konkuk University; Hyun Sook Yi, Konkuk University

  This study applies time-series clustering using dynamic time warping distance to identify response time patterns in low-stakes computerized adaptive testing. This analysis reveals behavioral disengagement types and supports real-time intervention strategies, offering practical implications for adaptive test design, response process analysis, and fairness in test delivery.

- **Exploring Slow Responses in Interactive Problem-Solving Tasks Using Sequential Process Analysis**
  Daniel Jerez, University of Alberta; Elisabetta Mazzullo, University of Alberta; Okan Bulut, University of Alberta

  We explored whether slow responses in an interactive problem-solving task in PISA 2012 are composed of a single or multiple subgroups. We applied full-path sequence analysis and clustering using process data. The analysed item has two clusters of students with slow responses who demonstrated distinct interactions with the testing environment.

- **Behavioral Patterns in Creative Thinking: A Latent Profile Analysis**
  Yaxin Dong, university of alberta; Xiaoxiao Liu, University of Alberta; Bin Tan, University of Alberta; Ying Cui, University of Alberta

  Latent profile analysis of students' creative thinking behaviors identified four distinct profiles: shifting-deliberate, moderate-deliberate, fluent-deliberate, and fast-intuitive. The ANOVA analyses show that these profiles differed in speed, time allocation, action rate, and performance. Multinomial logistic regression was then used to examine background and motivational predictors of profile membership.

- **Utilizing ML Models to Infer About Examinees' Cognitive Process on Coding Tasks**
  Kefan Yu; Min Li, University of Washington; Mo Zhang, Educational Testing Service; Chen Li, ETS; Amy Ko, University of Washington; Benjamin Zhou, University of Washington; Hongwen Guo, ETS Research Institute; Janet Jiang, University of Washington

  This study aims to investigate action logs on Python coding tasks to identify sequential patterns among undergraduate students varying in proficiency. Specifically, we generated sequential actions from events and editing types and then identified behavioral patterns across proficiency levels. Preliminary results show high-performing students approached tasks more systematically and strategically.

- **Using TFAM Sessions with Pre-service Teachers to Improve Classroom Assessment Practices**
  Lance Piantaggini, University of Massachusetts, Amherst

  This study introduces a novel framing of classroom assessment for 27 graduate students enrolled at a large New England university. Using a quasi-experimental design with a non-equivalent comparison group, initial findings show some increase in the understanding of, but not use of formative assessment.

- **Mitigating Early Aberrant Responses in Ability Estimation: The Weighted CAT Approach**
  Meng-Lin Li; Jyun-Hong Chen, National Cheng Kung University

  This study proposes a weighted CAT to address the bias that may arise by early aberrant responses. The simulation results reveal that, it matches conventional CAT in accuracy under normal conditions, and under conditions with aberrant responses, shows smaller bias, lower RMSE, faster recovery and improved item-selection quality.

- **Residual-Based DIF detection Framework for Operational Items in CAT: Evaluation and Application**
  Bomin Chung; Hwanggyu Lim, Inha University; Yongsang Lee, Inha University

  This study evaluates the residual-based DIF (RDIF) framework in operational items of CAT. Simulation studies compare RDIF with CATSIB across various conditions in terms of the detection accuracy, and an empirical CAT dataset illustrates its practical utility for ensuring fairness and accuracy in CAT.

- **Adaptive CUSUM Charts for Real-Time Detection of Item Parameter Drift in CD-CAT**
  Jing Huang, Purdue University; Hua-Hua Chang, Purdue University

  This study proposes adaptive CUSUMs for detecting item parameter drift within CD-CAT online calibration. Simulation and empirical studies compared adaptive CUSUMs, traditional CUSUMs, and LRT. The results indicated that adaptive CUSUMs provided more prudent signaling for smaller IPD size and showed superior performance in more challenging detection scenarios.

- **The Construct of a Multiple-Choice EFL Reading Test Through Process Modeling**
  Sao Bui, The University of Melbourne

  This study examines the cognitive processes underlying multiple-choice reading comprehension questions with evidence from a large-scale high-stakes English proficiency test in Vietnam, the VSTEP. Relevant cognitive processes are represented by text and item features, and their explanatory power for reading item difficulty is analyzed using the explanatory item response model.

- **Behind the Screen: Mode Effects in Reading and Math in Buenos Aires**
  Nicolas Buchbinder, University of Colorado Boulder

  I examine the effect of the mode of administration of standardized tests in reading and mathematics on student scores in primary schools in Buenos Aires, Argentina. I find moderate negative effects of taking the test on a computer, but only for public schools in reading, concentrated in male students.

- **Score Validity Across Modalities: TEI Parameter Use in Paper-Based Assessment**
  Andrea Hebert, Cognia

  This study examines whether using technology-enhanced item (TEI) parameters to score paper-equivalent (PE) items introduces bias in student ability estimates. Through independent calibration and simulation, we assess differences in scoring accuracy and classification. Findings highlight potential validity concerns for accommodated assessments and inform best practices in mixed-format testing programs.

- **Predicting Item Survival Using Data Augmentation**
  Mubarak Mojoyinola; Ye Lin, Ascend Learning

  Class imbalance presents significant challenges for predicting item survival in high-stakes assessments. We assessed four resampling methods with five machine learning algorithms on nursing examination items. The combination of Borderline-SMOTE and XGBoost yielded the highest performance (F1-score = 0.81), indicating that effective resampling strategies can substantially improve prediction accuracy.

- **Test-taking Effort Over Time: Implications for Value-Added Assessment in Higher Education**
  Mara McFadden, James Madison University; Joseph Kush, IXL Learning

  This study examines how students' test-taking effort changes throughout college and its influence on value-added estimates. Using longitudinal assessment data, we compare two different value-added estimate intervals. Findings inform how fluctuations in effort impact interpretations of institutional effectiveness, guiding more valid assessment and reporting practices in higher education.

- **Explaining Reading Literacy in High-Performing Asian Systems Using PIRLS 2021**
  Guanyu Chen, The University of British Columbia; Qi Qi, Iowa State University; Kuan-Yu Jin, Hong Kong Examinations and Assessment Authority; ZHEN WANG, Hong Kong Examinations and Assessment Authority; Xiaoming Xi, Hong Kong Examinations and Assessment Authority

  Using PIRLS 2021 data, this study applies hierarchical linear modeling to examine student, family, and school factors associated with reading achievement in Hong Kong and Singapore, guided by Walberg's theory of educational productivity. Results reveal distinct pathways to high performance and provide insights into improving educational productivity across contexts.

- **Rethinking Plausible Value Estimation in Large-Scale Assessments with Bayesian Model Averaging**
  Kjorte Harra, University of Wisconsin-Madison; David Kaplan, University of Wisconsin - Madison

  Plausible values (PVs) estimate student latent ability in large-scale assessments. While widely used to for achievement and secondary analyses, current PV methods currently ignore imputation model uncertainty, hindering results. We propose applying Bayesian Model Averaging (BMA) to PV estimation, streamlining implementation and improving predictive performance in secondary analyses.

- **Propensity Score Weighting and Effect Estimation with Sampling Weights and Plausible Values**
  Kaijie Liu, James Madision University; Joseph Kush, IXL Learning

  This study demonstrates various approaches for handling sampling weights and plausible values in propensity score estimation. Using PISA data, treatment effect estimates and their standard errors are compared across different approaches, informing bias and precision implications when not following the OECD (2009) guidelines while adding empirical evidence to the literature.

- **Toward a Multidimensional Framework: Missing Data and Item Characteristics in PISA Creativity**
  Deepak Sharma, IIM Ahemdabad

  This study investigates missing responses in PISA 2022 creative thinking items, analyzing how item attributes—question length, item order, input type, curriculum relevance, and multimedia presence—influence missingness. Findings inform item design for engagement optimization. Future work will develop a multidimensional creativity construct to enhance measurement validity beyond PISA's unidimensional approach.

- **Psychosocial Networks of Competition, Cooperation, Belonging, and Bullying among U.S. Adolescents**
  Kahee Han, University of Kansas; Hyeri Hong, California State University, Fresno

  Using PISA 2018 U.S. data (N = 4,175), this study applied Exploratory Graph Analysis to examine networks of competition, cooperation, belonging, and bullying. Results highlight belonging and cooperation as central, with bullying negatively linked to belonging. Network comparison revealed SES-based differences at the item level, suggesting targeted school climate interventions.

Advancing Process Data Analysis in Both High and Low Stake Assessments

Coordinated Paper Session
3:45 PM – 5:15 PM
Intercontinental Los Angeles Downtown, Floor 5 : Silver Lake A

Process data captured through human-computer interactions has become a vital source of evidence for enhancing validity and reliability in educational measurement. While widely used in low-stakes assessments to examine students' problem-solving strategies and behavioral patterns, its application in high-stakes contexts is rapidly expanding. This shift enables new functions for process data, including performance prediction, item design innovation, and the development of standardized analytical tools. This coordinated session features four novel studies that demonstrate the feasibility and implications of incorporating process data into digital assessments across both high- and low-stakes settings. The first paper introduces a two-step regression analysis on process data to predict student performance and identify unique behavioral profiles, including those of students with special learning needs. The second study optimizes response time thresholds with process data for computer-based case simulation tasks in high-stake assessments. The third paper analyzes online homework data to trace learning trajectories and inform formative assessment design. The final study presents a novel image-based method that transforms each respondent's process into visual profiles, enabling robust clustering of behavioral patterns through image processing and providing a generalized framework for analyzing sequential process data.

Discussant:
Alina von Davier (Duolingo)

Presentations:

1. **Predicting Exam Performance with Process Data in High-Stakes English Assessments**
   Liyuan Liu, Pearson; Blake Ashworth, Pearson; Hayley Dalton, Pearson; Qiwei He, Georgetown University; Barbara Donahue, Pearson

2. **Evaluating Time Reduction Effects in Simulation-Based Assessments Using Process Data**
   Chunyan Liu, National Board of Medical Examiners; Monica Cuddy, NBME; Qiwei He, Georgetown University; Wenli Ouyang, National Board of Medical Examiners; Cara Artman, National Board of Medical Examiners

3. **Using Browser Log Data to Measure Critical Online Reasoning**
   Jannick Illmann, DIPF | Leibniz Institute for Research and Information in Education; Daniel Baumartz, Goethe University, Germany; Philine Drake, DIPF; Carolin Hahnel, Ruhr Universität Bochum; Johannes Hartig, DIPF; Carmen Köhler, DIPF; Marcus Schrickel, DIPF; Alisa Stroganova, Centre for International Student Assessment (ZIB), Germany; Frank Goldhammer, DIPF | Leibniz Institute for Research and Information in Education

4. **Advancing Sequence Clustering with Image Processing in Interactive Assessments**
   Qiwei He, Georgetown University; Binhui Chen

Applications of AI in Psychometrics and Assessment

Individual Paper Session
3:45 PM – 5:15 PM
Intercontinental Los Angeles Downtown, Floor 5 : K-Town

Chair:
Hope Adegoke (University of North Carolina, Greensboro)

Discussant:
Nate Smith, ABIM

Presentations:

1. **Artificial Intelligence in Cheating Detection: Comparing Traditional and IRT-Based Approaches**
   Ismail Dilek

   This study compares traditional answer-copying indices and IRT-based residual methods for detecting cheating in simulated multiple-choice assessments. Results indicate that overlap indices effectively flag direct copying, while IRT residuals uncover broader anomalies. Exploratory AI clustering further reveals subtle collusion patterns, highlighting the value of multi-method, AI-enhanced frameworks for test security.

2. **Leveraging Transformer-based Models for Standard Alignment Study**
   Fang Peng, NWEA

   This study explores transformer-based retrieval models for educational standard alignment, comparing multiple models that rely solely on standards text with those incorporating item content. It further examines the degree to which fine-tuning contributes to improved performance.

3. **Variability and Carelessness in AI-Generated Psychometric Data: A Comparison with Human Responses**
   Elisabetta Mazzullo, University of Alberta; Okan Bulut, University of Alberta

   Careless responding is a common issue in psychometric tools. This study compared human and synthetic responses generated by GPT-3.5-Turbo to a validated mental health questionnaire. Our findings indicate that GPT-3.5-Turbo exhibited restricted variability in its responses. The model also displayed carelessness at rates comparable to those observed among human participants.

4. **Measuring the Alignment Between AI-Generated Stories and a K-2 Reading Curriculum**
   Qian Shen, University of Florida; Walter Leite, University of Florida; Jinnie Shin, University of Florida; Xiaomeng Xiong, University of Florida

   AI can support K-2 reading instruction by generating stories for reading practice. This study aimed to measure the alignment between AI-generated children's stories and a widely used reading curriculum. Stories were evaluated with a natural language processing (NLP)-based feature list. Additionally, latent class analysis was used to classify story suitability.

5. **Evaluating AI Prompts for Expert-Like Scoring Across Skill Complexity Levels**
   Merve Sarac, College Board; Joe Grochowalski, College Board; Lei Wan, College Board; Mark Klein, The College Board; Lauren Molin, College Board; Amy Hendrickson, The College Board

   This study evaluates whether generative AI can function as expert-like assessors of student essays. We compare training-material-oriented-prompts with expanded custom-prompts across multiple free-response tasks. Results show improved quadratic weighted kappa and stronger alignment with increasing skill complexity, suggesting AI prompts may capture deeper cognitive constructs with expert-like properties.

Classifications and Risk Assessment

Individual Paper Session
3:45 PM – 5:15 PM
Intercontinental Los Angeles Downtown, Floor 5 : Westwood

Chair:
Montserrat Valdivia Medinaceli (Curriculum Associates)

Discussant:
Heather Buzick (ACT)

Presentations:

1. **Classification Accuracy and Consistency Reference Tables for Practical Interpretation and Use**
   Adrienne Walker, Ascend Learning; Kari Hodge, Ascend Learning

   Guidance for evaluating classification accuracy (CA) and consistency (CC) observed in practice is scant. Using simulated data mimicking common IRT testing scenarios, we calculated benchmark values for CA and CC. These benchmarks can help researchers and practitioners interpret and evaluate the CA and CC values they observe in practice.

2. **Handling Missingness and Fairness in Group Cognitive Testing**
   Fatih Ozkan, Baylor University; Debi Torres, Baylor University; Uzeyir Ogurlu, Utah Valley University

   Using CogAT Forms 9–11, we profile domain-level missingness, reject MCAR for key subgroups, quantify disparities with risk differences, compare listwise deletion to multiple imputation of domain sums, and run MNAR tipping-point sensitivities. Results show practical, auditable steps that restore N without shifting gifted-rate conclusions

3. **Designing a Dyscalculia Screener for K-3**
   Sarah Quesen, WestEd; Sandra Pappas, Amplify

   This study developed and validated a dyscalculia risk flag for K-3 students (N=1,234) within mCLASS Math. Using individually administered WJIV subtests and multi-timepoint screener data, we derived empirically supported cutpoints through ROC analysis. Final thresholds identified 10.5% of students with persistent risk, providing defensible criteria for early identification.

4. **Evaluating the Quality of Multiple-Choice Items: A Nominal Response Model Application**
   Angel Arias, Carleton University

   This paper evaluates the quality of multiple-choice items on a listening test used to make high-stakes decisions by modelling polytomous data with the nominal response model. The paper puts particular emphasis on the interpretation of model parameters and the implications of this research for test development and validation research.

Collateral Resources to Support Use of the Revised Joint Standards

Organized Discussion
3:45 PM – 5:15 PM
Intercontinental Los Angeles Downtown, Floor 7: Hollywood Ballroom II

Session Description:
With a revised edition of the Standards for Educational and Psychological Testing currently in development, the testing field has an opportunity (and an obligation) to reflect on the potential value of supplementary resources that could accompany the release of the new edition. Such collateral materials can enhance the accessibility, interpretation, and application of the Standards among external audiences, including test users, test takers, and the general public. The panelists for this organized discussion are involved in the Standards revision and have championed the development of user-focused collateral resources. Panelists will share how the revision will incorporate material to support test users, discuss examples of resources developed to support the use of the 2014 Standards, and propose strategies for effective creation and dissemination of future supplemental resources. Audience members will be invited to consider how they, along with their respective organizations, might contribute meaningfully to this initiative.

Chair:
Jessica Jonson (Buros Center for Testing - University of Nebraska Lincoln)

Discussant:
Michael Walker (Human Resources Research Organization (HumRRO))

Presenter(s):
Laura Hamilton, National Center for the Improvement of Educational Assessment, Inc.; Kristen Huff, Curriculum Associates; Jessica Jonson, Buros Center for Testing - University of Nebraska Lincoln; Bradley McMillen, Wake County Public School Sustem; Deborah Bandalos, James Madison University

From Topic Models to LLMs: AI-Driven Applications in Educational Measurement

Coordinated Paper Session
3:45 PM – 5:15 PM
Intercontinental Los Angeles Downtown, Floor 7: Roosevelt B

The integration of artificial intelligence (AI) into educational assessment offers new opportunities to analyze complex learner data while addressing longstanding challenges in scoring, validity, and interpretability. This session brings together five studies that demonstrate systematic applications of natural language processing (NLP), including topic models and large language models (LLMs), in educational measurement. The first study introduces a transformer-based NLP pipeline for constructed-response scoring and reasoning analysis, showcasing how modern embeddings enhance both accuracy and interpretability. The second explores LLMs for qualitative coding of teacher interviews, comparing their performance with expert coders and examining strategies for effective prompting. The third study leverages multimodal data from computer-based assessments to derive item-level covariates using NLP and evaluates their predictability for psychometric models. The fourth presents a novel Topic Testlet Model, integrating topic modeling with testlet response theory to improve calibration and interpretability of constructed-response testlets. Finally, the fifth study develops a rater routing algorithm that combines text features with unfolding IRT models to improve the reliability of human-scored responses. Collectively, these contributions illustrate how AI can support scalable, validity-driven, and interpretable assessment design. The session concludes with a synthesis of findings and discussion of future directions for responsible AI integration in educational measurement.

Discussant:
Hong Jiao (University of Maryland)

Presentations:

1. **Transformer-Based NLP Pipeline for Constructed-Response Scoring and Reasoning**
   Constanza Mardones-Segovia, University of California, San Diego; Yaxuan Yang, University of Georgia; Cheng Tang, The University of Georgia; Jiawei Xiong, Curriculum Associates; Shiyu Wang, University of Georgia; Allan Cohen, University of Georgia

2. **Exploring Large Language Models' Capabilities in Qualitative Coding**
   Cheng Tang, The University of Georgia; Kun Wang, Rethink Learning Labs; Yaxuan Yang, University of Georgia; Shiyu Wang, University of Georgia; Allan Cohen, University of Georgia; Rachel Brown, The Pennsylvania State University; Chandra Orrill, Rethink Learning Labs

3. **Data-Driven Item Features Learning from Educational Assessment**
   Yaxuan Yang, University of Georgia; Jing Li; Eunji Lee, University of Georgia; Shiyu Wang, University of Georgia; George Engelhard, University of Georgia

4. **Calibrating Constructed Responses in Testlets with a Testlet Topic Model**
   Jiawei Xiong, Curriculum Associates; Cheng Tang, The University of Georgia; Huan Kuang, Florida State University

5. **A Rater Routing Algorithm: Integrating Text Analysis and Unfolding IRT Models**
   Jordan Wheeler, University of Nebraska - Lincoln; George Engelhard, University of Georgia

International Large Scale Assessment Research

Individual Paper Session
3:45 PM – 5:15 PM
Intercontinental Los Angeles Downtown, Floor 5 : Silver Lake B

Chair:
Mingfeng Xue (University of North Carolina Greensboro)

Discussant:
Paul Bailey (American Institutes For Research)

Presentations:

1. **Too Easy or Too Hard? Engagement in Adaptive Testing Using Process Data**
   Yuan-Ling Liaw, IEA Hamburg; Alec Kennedy, IEA Hamburg

   This study analyzes student engagement under TIMSS's group-adaptive design, an innovation in international large-scale assessment. Using process data and mixture modeling, we detect disengaged responses across countries. Findings show benefits of tailoring tests to student ability levels but also reveal limitations, indicating improvements are needed to ensure fairness across systems.

2. **Analyzing Trends in Students' Mathematics Attitudes and Gender-Achievement Gaps**
   Bethany Fishbein, Boston College; Ummugul Bezirhan, Boston College; Aurélie Lacroix, DEPP, French Ministry of Education; Franck Salles, DEPP, French Ministry of Education; Christian Kjeldsen, IEA

   This study examines mathematics attitudes and widening gender-achievement gaps observed over time, employing an IRT scaling approach to facilitate cross-country and cross-cohort comparisons. Findings reveal historical declines in attitudes, with gender-attitude gaps widening, paralleling achievement differences. The results raise questions about the evolving nature of affective constructs in analyzing achievement.

3. **Post-Selection Inference for Creative Thinking in PISA 2022 Korea and Colombia**
   Minjeong Rho; Andry Bustamante Barreto, Department of Education, Korea National University of Education; Jin-Eun Yoo, Department of Education, Korea National University of Education

   Using PISA 2022 student data from Korea and Colombia, this study applies LASSO and post-selection inference to identify statistically significant predictors of creative thinking. Findings reveal both shared and country-specific factors, highlighting methodological advances and implications for equity-focused educational policies and large-scale assessment research.

4. **ILSAmerge and ILSAstats: Two new R packages for international large-scale assessments**
   Andrés Christiansen, IEA Hamburg

   We present two new R packages, ILSAmerge and ILSAstats, that address challenges in analyzing International Large-Scale Assessment (ILSA) data. ILSAmerge simplifies downloading and merging data, while ILSAstats provides accurate statistical analysis that accounts for plausible values and replicate weights.

5. **Lost in Translation: Quantifying Semantic Drift in Multilingual Translations**
   Ummugul Bezirhan, Boston College; Ji Yoon Jung, Boston College; Matthias von Davier, Boston College

   This study examines semantic drift introduced by translating multilingual student responses into English for automated scoring. Using LaBSE and multilingual-E5 embeddings, we compare original and translated responses with global, local and pairwise metrics. Results show that LaBSE preserves semantic geometry more reliably while E5 shows better pairwise similarity.

Methodological Considerations for DIF Detection

Individual Paper Session
3:45 PM – 5:15 PM
Intercontinental Los Angeles Downtown, Floor 5 : Boyle Heights

Chair:
Hyeonjoo Oh (Riverside Insights) and Hao Zhou

Presentations:

1. **Detecting Uniform DIF with Small Group Sizes in Certification Exams**
   Sarah Pirani, American Osteopathic Association; Lidia Martinez, American Osteopathic Association

   Most DIF research assumes large samples, yet credentialing programs often involve very small groups. This simulation study evaluates three common DIF methods under realistic certification conditions, focusing on detecting uniform DIF when subgroup sizes are limited to 25–50 examinees.

2. **Beyond the Score: Exploring DIF Detection Methods on a Medical Board Exam**
   Autumn Wild, James Madison University; Caroline Prendergast, American Board of Surgery; Carol Barry, American Board of Surgery

   This study compared the feasibility of three DIF detection methods – Rasch DIF, MNLFA, and Third Generation DIF – applied to medical board exam data. The Rasch method demonstrated superiority in implementation, whereas the other methods faced practical issues. Implications for certification and licensure examinations are discussed.

3. **Regularized DIF Detection in Multidimensional Graded Response Models**
   Weicong Lyu, University of Macau; Yijun Cheng; Ruoyi Zhu, University of Washington; Chun Wang, University of Washington; Gongjun Xu, University of Michigan

   We propose two regularization methods for detecting differential item functioning (DIF) in multidimensional graded response models with respect to respondent-level covariates. By modeling both the means and covariance structures of latent traits, the proposed approaches improve flexibility and accuracy in identifying DIF for continuous and discrete covariates.

Methodological Issues in DCM

Individual Paper Session
3:45 PM – 5:15 PM
Intercontinental Los Angeles Downtown, Floor 5 : Ladera Heights

Chair:
Chia-Yi Chiu (Teachers College, Columbia University)

Discussant:
Yale Quan (University of Washington)

Presentations:

1. **When is a Generically Identified Q-Matrix Truly Identified?**
   Jimmy de la Torre, The University of Hong Kong; Wenchao Ma, University of Minnesota; Sangbeak Ye, Florida Atlantic University

   The conditions for the generic Q-matrix identifiability involving general restricted latent class models exist. However, this paper shows that, for a given set of CDMs, identifiability may or may not hold depending on how the CDMs are configured. As an alternative, an empirical procedure for determining Q-matrix identifiability is proposed.

2. **A General Approach to Q-Matrix Refinement in Cognitive Diagnosis**
   Yingqi Huan, Teachers College, Columbia University; Chia-Yi Chiu, Teachers College, Columbia University

   The Q-matrix specifies item-attribute associations in cognitive diagnosis, yet misspecification undermines test validity and classification accuracy. This study presents a General Q-matrix Refinement (GQR) method using residual sum of squares (RSS) as loss function. The method is theoretically grounded, applicable to various CDMs, and its performance is evaluated empirically.

3. **Using Causal Discovery to Create Attribute Maps from Cognitive Diagnostic Models**
   He Ren, University of Washington; Chun Wang, University of Washington

   We propose a hybrid approach combining causal discovery and ordering theory to uncover attribute relationships in cognitive diagnosis. Viewing hierarchies as a special causal structure, we derive the attribute map capturing strict hierarchies and general relations. Simulations show robust recovery of these maps under diverse conditions.

4. **An Area-Level Small Area Bayesian Dynamic Borrowing Model**
   Sinan Yavuz, Amazon

   This paper presents an area-level Small Area Bayesian Dynamic Borrowing (SABDB) model as a computationally efficient alternative to unit-level approaches. K-means clustering is used to group similar states, providing a stable structure for information borrowing. The model yields more accurate subgroup achievement estimates in large-scale assessments.

Practical Issues and Innovative Solutions in Standard Setting

Individual Paper Session
3:45 PM – 5:15 PM
Intercontinental Los Angeles Downtown, Floor 6: Majestic

Chair:
Eric Asare (Old Dominion University)

Discussant:
Adam Wyse (Renaissance)

Presentations:

1. **Reinforcement Learning for Automated Bookmarking Standard Setting to Reduce Operational Burden**
   Igor Himelfarb; James Zoucha; Bruce Shotts, NBCE

   This study presents a reinforcement learning framework to simulate SME bookmarking in standard setting. Using interpretable statistical metrics, the approach closely replicated historical SME cut scores, achieving high correlations and low error rates. Results highlight potential cost savings, scalability, and validity preservation for operational testing programs.

2. **Exploring the Effect of Feedback on Standard Setting Ratings: A Mixed-Methods**
   Chansoon (Danielle) Lee, American Board of Internal Medicine; Kelly Rewley

   Using a mixed-methods approach, this study explores the effect of feedback on content experts' standard setting ratings. Preliminary findings from the quantitative analyses suggest that providing feedback (e.g., projected pass rates) significantly impacts ratings in ways that may undermine the content-based nature of the process.

3. **Computing Passing Scores from Matrixed Standard Setting Designs: A Comparison of Methods**
   Meng Fan, HUMRRO; Michael Walker, HumRRO

   This research explored ways to estimate from matrixed standard setting designs what would have resulted had all panelists rated all items. We found that in the absence of rater-by-item interactions, straight averaging without adjustment provided the most faithful results. Results could differ in the presence of interactions.

4. **Applying Latent Class Analysis to Validate Performance Standards on Large-scale Educational Assessment**
   Kyle Du, CUNY Graduate Center; Howard Everson, City University of New York; Jay Verkuilen, CUNY Graduate Center

   Latent class analysis (LCA) was used to validate proficiency standards on a statewide standardized math test. Results indicate the number and meaning of the LCA derived proficiency categories align with those of the subject matter experts. LCA methods, however, may produce more stringent cutscore boundaries than the expert panelists.

Understanding Assessment Contexts and Social-Emotional Competencies

Individual Paper Session
3:45 PM – 5:15 PM
Intercontinental Los Angeles Downtown, Floor 7: Roosevelt A

Chair:
Fernando Mena Serrano

Discussant:
Christina Cipriano (Yale University) and Fernando Mena (University of Massachusetts Amherst)

Presentations:

1. **Growth in Social Emotional Competence Can Mitigate Chronic Absenteeism Risk**

   Evelyn Johnson, Riverside Insights; Emily Taylor, Riverside Insights; Jennifer Robitaille, Riverside Insights

   We demonstrate how a one-minute screener of social emotional competence (SEC) can help address chronic absenteeism. Logistic regression showed students with a need for instruction were 3.3 times as likely to be chronically absent. Propensity score matching examined whether substantial growth on SEC mitigated risk and confirmed significantly lower absenteeism.

2. **Strengthening Social-Emotional Skills to Reduce Chronic Absenteeism**
   Evelyn Johnson, Riverside Insights; Emily Taylor, Riverside Insights; Jennifer Robitaille, Riverside Insights

   This study examines whether students' social-emotional competence (SEC) predicts chronic absenteeism in a large high school sample. Logistic regression and generalized structural equation modeling showed that intrapersonal and decision-making skills reduced absenteeism risk, highlighting SEC as a potential solution to help reduce chronic absenteeism.

3. **Thanks, I Hate It: Addressing Negative User Experience of Forced-Choice Assessment**
   Kevin Williams, ETS; Diego Figueiras, ETS; Katrina Roohr, ETS; Guangming Ling, ETS; Jianbin Fu, ETS

   We conducted two studies to evaluate a new assessment format called hybrid multidimensional forced-choice/ Likert (HMFCL), which combines multidimensional forced-choice (MFC) and Likert. Results suggested that HMFCL demonstrated superior user experience (accuracy, fairness, enjoyability, etc.) and similar or superior fake-resistance (i.e., score inflation) relative to MFC and Likert, respectively.

4. **When You Test Matters: Impact of Time of Test on Student Performance**
   Ian Campbell, Cambium Assessment

   Accountability testing in state assessments must balance standardizing conditions across the entire state with logistical realities. To do this, many states allow schools to schedule testing opportunities within a window of time. This study analyzes the impact of when in the window tests occur on performance and testing behavior.

5. **In-program Formative Situational Judgment Test : Competency Framework Analysis and Pilot Study Findings**
   Gabriel Sitarenios, Acuity Insights; Rodica Ivan, Acuity Insights; Muhammad Iqbal, Acuity Insights; Cole Walsh, Acuity Insights; Josh Moskowitz, Acuity Insights

   Academic programs emphasize supporting students to develop professional competencies. We constructed a formative open-response situational judgment test synthesizing 95 competencies across multiple frameworks into four domains: intrapersonal, interpersonal, social/ethical responsibility, and critical thinking. Piloted with 350 students across three programs, psychometric analyses supported its use as a formative tool.

State & Local Assessment Leaders (SALAL) SIGIMIE

Meeting
4:15 PM – 5:15 PM
Intercontinental Los Angeles Downtown, Floor 6: Royal

This SIGIMIE is designed for state and local assessment leaders from around the country. It's built around the concept of building a community of practitioners who are "doing assessment" on a daily basis in state education agencies as well as large school districts.

This community spends time discussing current affairs in educational assessment and working together to support each other in navigating the ever changing world of student assessment. Individuals that may be independent contractors or work for an assessment vendor are welcome to join, but may be excused for certain conversations. Our group focuses on monthly meetings, webinars, conference sessions, peer support, and collaboration. Please reach out for more information.

Large-Scale Assessments SIGIMIE

Meeting
5:30 PM – 6:30 PM
Intercontinental Los Angeles Downtown, Floor 6: Royal

Standardized assessments and surveys play an important role in shaping educational decisions around the globe and at multiple levels – from classrooms to local, regional, and national governance.

The Large-Scale Assessment (LSA) SIGIMIE is focused on how results from assessments including licensure exams, K-12 accountability assessments, international assessments, and others can be better leveraged to inform thoughtful, equitable, and evidence-based decision-making. We do this by looking at three aspects of LSAs: instrumentalization, test use/misuse, and concepts of fairness.

We are particularly interested in how the LSA SIGIMIE can inform assessment policy. This requires understanding how a range of individuals engage with and respond to assessments. This SIGIMIE aims to identify ways to improve design and development practices, support appropriate interpretation of results, and recognize limits of test use through understanding the strengths and limitations of using such tools with a variety of populations.

Joint NCME/AERA Division D Welcome Reception

Social
7:15 PM – 8:45 PM
Intercontinental Los Angeles Downtown, Floor 5 : Wilshire Grand Ballroom Prefunction

Joint NCME/AERA Graduate Student Reception

Social
8:30 PM – 10:00 PM
Intercontinental Los Angeles Downtown, Floor 7: Hollywood Ballroom Patio

FULL SCHEDULE
FRIDAY, APRIL 10

Coffee & Bagels
Social
7:30 AM – 9:00 AM
Intercontinental Los Angeles Downtown, Floor 5 : Wilshire Grand Prefunction

Hot coffee & tea, bagels, croissants and danishes will be available.

NCMEntoring Program
Meeting
7:30 AM – 5:00 PM
Intercontinental Los Angeles Downtown, Floor 6: Roxy

Launched at the 2016 annual meeting, the NCMEntoring Program aims to support the transition of graduate student members and recent graduate members from their graduate programs to professional careers. Early professionals (mentees) are paired with members (mentors) experienced in NCME-related fields: psychometrics, assessment, certification, evaluation, and other aspects of educational measurement.

This experience offers mentees the opportunity to explore possible career paths and/or research interests and for mentors to support the development of potential colleagues and contribute to the field. Each year, over 100 NCME members participate in the NCMEntoring Program and participant feedback has been positive. The Program hopes to cultivate long-term relationships between mentors and mentees.

Low Sensory Room
Meeting
7:30 AM – 6:00 PM
Intercontinental Los Angeles Downtown, Floor 6: Jade

Attendees who need areas that are quiet with reduced light and noise may take comfort in the low-sensory room.

Mother's Room
Meeting
7:30 AM – 6:00 PM
Intercontinental Los Angeles Downtown, Floor 5 : Glassell Park

Private mother's room for nursing.

NCME Business Meeting/Presidential Address
Invited Session
7:30 AM – 9:00 AM
Intercontinental Los Angeles Downtown, Floor 5 : Wilshire Grand Ballroom III

Quiet Room
Meeting
8:00 AM – 7:00 PM
Intercontinental Los Angeles Downtown, Floor 6: Novelty

During the Annual Meeting, attendees who desire a quiet place to relax or prepare for a presentation may visit quiet rooms available 8:00 am-7:00pm daily.

Individual eBoards: DIF and Dimensionality

Electronic Board Session
9:45 AM – 10:45 AM
Intercontinental Los Angeles Downtown, Floor 5 : Hancock Park

Presentations:

- **Detecting Polytomous DIF: Comparing Two Recursive Partitioning Approaches (eBoard 1)**
  Nana Amma Asamoah, University of Arkansas; Ronna Turner, University of Arkansas

  This study evaluates two tree-based methods (global and item-focused) for detecting differential item functioning in polytomous items. Simulations involving diverse measurement conditions previously untested are carried out, and results reveal method-specific strengths and limitations, offering critical insights for practitioners and researchers in selecting the appropriate method.

- **Comparing Measurement Invariance Methods in Clinical Assessment (eBoard 2)**
  Hsin-Ro Wei, Riverside Insights; Hyeonjoo Oh, Riverside Insights; Tong Wu

  This study evaluates multi-group confirmatory factor analysis (MGCFA), exploratory graph analysis (EGA), and structural equation model trees (SEMTree) applied to clinical assessment data. Results supported configural and metric invariance across gender but revealed limited scalar invariance, indicating valid relational cross-gender comparisons but caution in interpreting factors mean differences.

- **Decision Tree Detects Sex and Race DIF in YRBSS Sexual Behavior Items (eBoard 3)**
  Sena Gunes, MEF University; Onur Ramazan, Washington State University

  We employed Decision Tree to detect DIF in sexual behavior items in the YRBSS. Results demonstrated that seven items functioned equivalently across sex and racial/ethnic groups although one item (“sex of sexual contacts”) showed significant sex-based measurement bias. Findings could help ensure equitable public health surveillance for diverse adolescent populations.

- **Differential Item Functioning Detection Methods in Mixed-Format Assessments Using Propensity Score Weighting (eBoard 4)**
  Minju Hong, Chung-Ang University; Hyesun You, University of Iowa

  For test fairness, we examine logistic regression (LR) using propensity score weighting. We prove LR’s flexibility to detect differential item functioning (DIF) items under various test conditions by incorporating dichotomous and polytomous items, uniform and nonuniform DIF, and grouped variables and covariates. Findings will guide best practices in psychometric applications.

- **Examining the Stability of DIF Classifications in Mixed-Format Tests with Unbalanced Subgroups (eBoard 5)**
  Adrienne Walker, Ascend Learning; Mubarak Mojoyinola

  Sample size imbalance is a common challenge in practical DIF analyses. We investigated DIF detection stability using bootstrapping versus single-draw balanced resampling and the original (unbalanced) sample. DIF was observed for four items and DIF classifications varied across the original and balanced sample techniques. Practical DIF detection implications are discussed.

- **Consensus Among DIF Effect Size Measures in Dichotomous Item Response Theory Models (eBoard 6)**
  Austin Wyman, University of Notre Dame

  The size of DIF is an important question but it is rarely reported because there are many effect sizes with limited interpretation guidelines. The present study evaluates the consensus among DIF effect size measures in dichotomous IRT models and develops novel interpretation recommendations to improve the accessibility of DIF studies.

- **Collecting Validity Evidence for the IXL LevelUp Math Assessment (eBoard 7)**
  Xiaozhu An, IXL Learning; Bozhidar Bashkov, IXL Learning

  We gathered validity evidence for the IXL LevelUp math assessment, a recently released measure embedded within IXL's widely used online learning platform. Findings supported its internal structure and multigroup measurement invariance across key student subgroups, which provides strong validity evidence of fairness and supports score interpretations across diverse student populations.

- **The Effect of Model Misspecification on Rasch Tree DIF Detection (eBoard 8)**
  Andrew Krist; Sanford Student, University of Delaware; Stefanie Wind, The University of Alabama

  This study investigates how common model misspecifications affect Rasch Tree differential item functioning detection. Using simulations, we examine the stability of node selection and contrast estimates when 2PL and 3PL features (discrimination and guessing) are introduced. Findings highlight the robustness and limitations of Rasch Trees under non-Rasch conditions.

- **Third-Generation DIF Analysis of PISA 2022 Well-Being Data (eBoard 9)**
  Weiran Li, The University of British Columbia

  This study applies Zumbo's third-generation DIF framework to PISA 2022 Canadian well-being data. Using logistic regression DIF analyses, we examine subgroup differences by gender, language, and immigration status. Findings highlight language-related DIF and underscore the importance of culturally responsive approaches for validity and fairness in international large-scale assessments.

- **Comparing Differential Distractor Functioning Detectors: Evidence from Simulation and Empirical Analyses (eBoard 10)**
  Keyu Chen, NCBE; Xiaoli Jiang; Mengyao Zhang, National Conference of Bar Examiners

  This study compares four promising methods for detecting differential distractor functioning (DDF) in multiple-choice items. Both simulated and empirical data are used to assess the performance of these methods. Results offer useful insights for optimizing DDF detection in practice to improve exam quality and fairness.

- **Evaluating the Psychometric Properties of the PLCA-R Using Rasch Analysis: A Rasch Partial Credit Model Approach (eBoard 11)**
  Justice Dadzie, The University of Alabama; Ruth Annan-Brew, University of Cape Coast, Ghana; Daniel Oyeniran, The University of Alabama; Christoper Ocheni, The University of Alabama, Tuscaloosa; Frank Oppong, Ohio University

  This study examined the Shared and Supportive Leadership (SSL) subscale of the PLCA-R using the Rasch Partial Credit Model. Findings showed strong reliability and mostly acceptable fit, though a few items misfit expectations. Differential Item Functioning indicated minimal subgroup bias. Overall, the subscale appears psychometrically sound but needs targeted refinement.

- **Comparing Bayesian Approximate Invariance Methods for Establishing Measurement Equivalence (eBoard 12)**
  Yichi Zhang, Georgia Institute of Technology

  Several Bayesian approximate invariance methods were proposed to adjust for partial measurement invariance. This study conducts a Monte Carlo simulation to compare approaches in terms of empirical test size and power, offering guidance on how Bayesian techniques can improve the practical evaluation of measurement equivalence in applied research.

- **Country-Level Differences in Response Distributions on the Student Bullying Scale (eBoard 13)**
  Henrietta Tettey-Tawiah; Ronna Turner, University of Arkansas

  This study examined items on a student bullying scale, used on an international academic assessment for fourth grade students, to compare how items function for students in highest vs. lowest performing countries and across language versions. The global tree approach indicated significant differences based on country performance level and language.

- **Bayesian Detection of Differential Item Functioning Using Posterior Predictive Model Checking (eBoard 14)**
  Sean Joo, University of Kansas; Philseok Lee, George Mason University

  This study introduces a Bayesian approach for detecting differential item functioning (DIF) through posterior predictive model checking (PPMC). Simulation results across DIF types, sample sizes, and item parameters demonstrate superior Type I error control and strong power for uniform DIF. Practical guidelines and real-data analyses highlight implications for large-scale assessments.

- **Comparing Scale Dimensionality Using Human Responses and Text Embeddings (eBoard 15)**
  Chia-Lin Tsai, University of Northern Colorado; Lisa Flores, University of Missouri-Columbia; Rachel Navarro, University of North Dakota ; Pat Garriott, University of Denver; Han Na Suh, University of Hawaii at Hilo; Bo Hyun Lee, Indiana University

  This study compares dimensions derived from participant perceptions, reflected in human responses, with those derived from the semantic meaning of items, captured through text embeddings. Using exploratory graph analysis, we analyzed data from both approaches. Our findings provide insights into evaluating and interpreting dimensions during the scale development process.

- **Rotation Matters: Sample Size, Test Length, and ESEM Performance Across Conditions (eBoard 16)**
  Talal Alzabidi, University of North Carolina Greensboro

  We evaluate ESEM rotations (Geomin ε=.01/.50, Quartimin, Target) across N=300, 500, 1000, and 12 versus 20 items using Monte Carlo. Geomin/Quartimin yields a cleaner, simpler structure than Target. Accuracy improves with N; for 20 items, Geomin ε=.50 minimizes RMSE. ε=.01 is sparser but less accurate. Recommendations are provided for practice.

- **Contextualizing Belonging in Higher Education: Instrument Development and Validation (eBoard 17)**
  Catherina Villafuerte, University of Connecticut

  This presentation validates a higher education belonging instrument built from institutional sites and mechanisms. Using ordinal confirmatory factor analysis and multi group invariance with DIF screening, we establish structure, reliability, and comparability, examine external relations with climate and mentoring, and provide score interpretation guidance and documentation for responsible institutional use.

- **A Text-Augmented CFA Model with a Novel Gibbs Sampling Algorithm (eBoard 18)**
  David Arthur, University of Washington; Yuxiao Zhang, Purdue University

  Confirmatory Factor Analysis (CFA) is a popular tool for measuring latent traits. However, CFA does not readily allow one to incorporate additional information about latent traits from text data. We introduce a novel text-augmented CFA model along with an efficient Gibbs sampling algorithm for estimation of this model.

- **Revisiting Reading Profiles: An Updated Approach to Understanding Skill Integration (eBoard 19)**
  Jonathan Weeks, Stanford University; John Sabatini, Institute for Intelligent Systems, University of Memphis

  This study extends prior examinations of reading skill profiles for over 165,000 students using a vector-based approach that incorporates KL divergence as an objective criterion to differentiate between levels of skill integration. Distinct profiles are then identified for each level, thus reducing the likelihood of identifying flat score patterns.

- **Examining the Association between Psychological Factors and Students' Mathematics Performance (eBoard 20)**
  Jeongmin Ji, University of Iowa; Hyeri Hong, California State University, Fresno

  This study examines the relationship between psychological factors and math achievement using social network analysis of the United States students in the PISA 2022 dataset. Findings show nodes within each dimension are connected: math efficacy relates directly, while teacher support relates indirectly, to math achievement.

- **Fit or Fiction? When Unidimensional IRT Models Fall Short (eBoard 21)**
  Ketan ., University of Massachusetts, Amherst; Craig Wells, University of Massachusetts Amherst

  This study examines when unidimensional IRT models are "good enough" for large-scale assessments. Using data from an international numeracy test and Monte Carlo simulations, we evaluate misfit, score recovery, and classification accuracy across uni- and multidimensional models. Findings inform valid and policy-relevant score reporting in global education contexts.

- **Gender Differences in Laparoscopic Simulator Performance: A Metric-Based Comparison (eBoard 22)**
  Noa Gazit; Gilad Ben-gal, Hebrew University of Jerusalem; Ron Eliashar, Hebrew University of Jerusalem

  This study examined gender differences in laparoscopic simulator performance among interns, residents, and experts. Men outperformed women at early training stages, but not among experts. Differences appeared only in specific metrics (success rate, time, path length) and not others, demonstrating that gender effects depend on the metric assessed.

- **Comparing Methods for Detecting IRT Parameter Drift in Common Item Equating (eBoard 23)**
  Blaine Pedersen

  Assessment programs often employ a common items non-equivalent groups equating design. The presence of items with parameter drift may decrease equating accuracy. This research evaluates the extent to which popular approaches, yet to be directly compared, for detecting and removing items with parameter drift from equating improve equating accuracy.

Advancing Conversation-Based Assessment with Large Language Models

Coordinated Paper Session
9:45 AM – 11:15 AM
Intercontinental Los Angeles Downtown, Floor 7: Roosevelt B

Conversation-based assessment (CBA) provides an innovative approach to testing as it builds on dialogues to capture what learners know and how they reason. With large language models (LLMs), it is now possible to design, deliver, and score CBAs at scale, creating new opportunities for authentic, engaging, and adaptive assessment. This symposium brings together five empirical studies from Canada, Germany, and the United States that advance research on CBA's cognitive, motivational, instructional, and psychometric dimensions. Presentation 1 compares multiple-choice, constructed-response, and CBA tasks using Bloom's taxonomy, assessing whether LLM-enabled dialogue can better elicit higher-order thinking. Presentation 2 contrasts CBA with multiple-choice assessment (MCA) in a controlled experiment, showing longer engagement and improved test-taking experiences in CBA over MCA. Presentation 3 presents a classroom experiment, testing whether the feedback structure in CBA vs. MCA with feedback improves not only engagement but also retention and transfer of learning. Presentation 4 explores learner modeling, using LLMs to infer knowledge components from CBA dialogues and comparing results with expert judgments as a step towards scalable personalization. Presentation 5 examines human–machine score disagreements, showing that experts tend to draw on implicit information and added context, offering insights to improve LLM scoring prompts and rubrics.

Discussant:
Arthur Graesser (University of Memphis)

Presentations:

1. **Comparing Perceived Cognitive Demand Across Different Item Formats**
   Okan Bulut, University of Alberta; Seyma Yildirim-Erbasli, Concordia University of Edmonton; Bin Tan, University of Alberta

2. **How LLM-Driven Conversation-Based Assessment Enhances Engagement and Test-Taking Experience Beyond Multiple-Choice**
   Marlit Lindner, IPN - Leibniz Institute for Science and Mathematics Education; Lauritz Schewior, IPN - Leibniz-Institute for Science and Mathematics Education; Guher Gorgun, Leibniz Institute for Science and Mathematics Education

3. **Learning Through Dialogue? A Classroom Experiment on LLM-Driven Conversation-Based Assessment versus Multiple-Choice**
   Lauritz Schewior, IPN - Leibniz-Institute for Science and Mathematics Education; Marlit Lindner, IPN - Leibniz Institute for Science and Mathematics Education

4. **Exploring Causes for Disagreement in Human and Machine Scores of Conversation-Based Assessments**
   Diego Zapata-Rivera, ETS; Carol Forsyth, ETS; Liang Zhang, University of Georgia

5. **Toward an Integrated Learner Model in a Conversation-Based Assessment System**
   Carol Forsyth, ETS; Diego Zapata-Rivera, ETS; Liang Zhang, University of Georgia; Jessica Andrews-Todd, ETS

DIF Research Applications

Individual Paper Session
9:45 AM – 11:15 AM
Intercontinental Los Angeles Downtown, Floor 5 : Boyle Heights

Chair:
Ann Arthur (ACT Education Corp.)

Discussant:
Yunhang Yin (University of South Carolina)

Presentations:

1. **Evaluating Longitudinal Measurement Invariance of an SEL Assessment Using Graded Response Models**
   Tianying Sun; Pui-Wa Lei, The Pennsylvania State University; James DiPerna, The Pennsylvania State University; Yue Tang, The Pennsylvania State University; Susan Hart, The Pennsylvania State University; Kyle Husmann, The Pennsylvania State University

   This study examined measurement invariance of a teacher-rated SEL assessment using graded response models. DIF analysis revealed minimal item/test-level bias. Subsequent group mean comparisons and theta correlations between full and partial invariance models indicated negligible consequences of DIF. Findings support temporal score stability and suggest item selection for assessment refinement.:

2. **Harnessing Constructed-response Inputs to Interpret Differential Item Functioning**
   Fusun Sahin, Curriculum Associates; Sebastian Moncaleano-Wallrich

   This study explores how students' inputs to constructed-response items can support interpreting Differential Item Functioning (DIF) between transadapted mathematics items. By analyzing response uniqueness, frequency rankings, and patterns across English and Spanish versions, we offer scalable, data-driven methods to aid expert review and identify potential bias.

3. **Generalized Methods for Intersectional Differential Item Functioning: Parsimony in Multiple Comparisons**
   Tony Albano, University of California - Davis; Oscar Rios, University of California - Davis; Leyna Kataoka, University of California, Davis; Xiaochen Xu, University of California, Davis; Demi Robinson, University of California, Davis

   Can generalized methods improve DIF analysis with intersectional groups? In our study, we compare simple vs generalized Mantel-Haenszel and logistic regression using simulated data. We evaluate new significance thresholds and show how generalized methods with omnibus tests are more parsimonious and accurate with a large numbers of focal groups.

4. **Innovative-Adaptive Model of TTCT Figural: Invariance Across Ethnicity, Not Gender**
   Yoojoong Kim, University of Georgia; Denis Dumas, University of Georgia; Selcuk Acar, University of North Texas; Peter Organisciak, University of Denver

   TTCT-Figural is the most influential creative thinking measure in U.S. schools, especially for gifted identification. In our dataset, the Innovative-Adaptive two-factor model showed the best fit to elementary students' data, and was invariant across ethnicities, though not across genders. A more nuanced interpretation of TTCT scores across genders is needed.

Designing an Assessment System to Support Meaningful Teaching and Learning

Organized Discussion
9:45 AM – 11:15 AM
Intercontinental Los Angeles Downtown, Floor 5 : K-Town

The California Science Test (CAST) was first administered operationally in 2019. Developed to align with the state's Next Generation Science Standards (CA NGSS), the test's design was intended to assess three-dimensional science teaching and learning. Since its introduction, a great deal has been learned about how to teach and assess the CA NGSS. Simultaneously, massive changes have occurred in the technologies available to assist those processes (e.g., generative AI). In combination, these factors argue for improving CAST to model and incentivize even more meaningful teaching and learning, while still serving its primary purpose as a federally mandated accountability measure. This organized discussion will explore how the nation's most populous state, California, proposes to re-envision its science assessment to achieve ambitious measurement and educational impact goals. Representatives from the State Board of Education, the State Department of Education, the Los Angeles County Office of Education, and state contractors will share their perspectives on the CAST vision and its implementation.

Chair:
Randy E Bennett, ETS

Discussant:
Randy E Bennett, ETS

Presenter(s):
Randy Bennett, Assessment Innovation Matters; Aneesha Badrinarayan, Education First; Linda Darling-Hammond, California State Board of Education; Ingrid Roberson, California State Department of Education; Deborah Atwell, Los Angeles County Office of Education; Hilary Persky, ETS Research Institute

Estimation with Complex Distributions and Data Structures

Individual Paper Session
9:45 AM – 11:15 AM
Intercontinental Los Angeles Downtown, Floor 7: Roosevelt A

Chair:
He Ren (University of Washington)

Discussant:
Wei-Chia Su (National Sun Yat-sen University)

Presentations:

1. **Just a Zero? Misspecification and Identifiability in Zero-Inflated Poisson IRT Models**
   Allison Ames

   Zero-Inflated Poisson IRT (ZIP-IRT) models count-based assessment data with excess zeros. Two concerns have not been addressed: (1) the sensitivity of ZIP-IRT to misclassification of zero-generating processes, and (2) the identifiability of the zero-inflation parameter. This simulation evaluates ZIP-IRT under model misspecification and latent dependency between ability and structural zeros.

2. **Integrating Pairwise Dissimilarity Data to MGGUM to Improve Estimation with Smaller Samples**
   Zhaoyu Wang, Georgia Institute of Technology

   This study improves the Multidimensional Generalized Graded Unfolding Model (MGGUM) by integrating Multidimensional Scaling (MDS) to improve parameter estimates and reduce sample size requirements. A simulation study will evaluate model performance under varied design factors. Findings are expected to advance practical applications of unfolding multidimensions IRT models in measurement research.

3. **Stacking Multiple Imputation for Missing Data in Item Response Theory**
   Yon Soo Suh, NWEA within HMH; Minho Lee, University of California - Los Angeles

   This study introduces a computationally efficient stacking multiple imputation (MI) approach for item response theory models with missing data. Unlike traditional MI, stacking combines imputed datasets into one weighted analysis, enabling point estimation and limited-information goodness-of-fit testing. Simulation and empirical studies demonstrate improved efficiency, accuracy, and model fit evaluation.

4. **Likelihood-based Estimation of Model-derived Oral Reading Fluency under Dispersed Count-time Data**
   Xin Qiao, University of South Florida; Emma Evudottir, University of South Florida; Cornelis Potgieter, Texas Christian University; Akihito Kamata

   It is crucial to make accurate inferences on oral reading fluency (ORF) scores for identifying at-risk students in various educational settings. We provide derivations and demonstrations of likelihood-based estimators of ORF scores when the reading count data has potential overdispersion.

5. **Robust Estimation of Latent Traits with the Multidimensional Graded Response Model**
   Audrey Filonczuk; Ying Cheng, University of Notre Dame

   A robust estimator for latent traits is proposed for the multidimensional graded response model (MGRM; Muraki & Carlson, 1995). The estimator addresses aberrant responses through systematic downweighting, reducing their influence on ability estimation. Simulation results demonstrate reduced bias across various test conditions, revealing improved accuracy amidst multidimensional, polytomous data.

Expanding the Scope of Test-Taking Engagement Research

Coordinated Paper Session
9:45 AM – 11:15 AM
Intercontinental Los Angeles Downtown, Floor 5 : Ladera Heights

This session presents new, innovative research that expands the way in which test-taking engagement is researched to understand how we design for, detect, and promote test-taking engagement. Leading researchers from Canada, Europe, and the United States will present their test-taking engagement findings. Presentation 1 furthers our understanding of the impact of item difficulty on test-taking engagement by creating a personalized item difficulty metric rather than a general metric. Presentation 2 expands the use of process data to identify test-taking engagement by exploring action sequences during interactions with a technology-enhanced item. Presentation 3 investigates test-taking engagement in the context of AI-based adaptive dialogue tasks to determine if they are more engaging than traditional assessment formats and identify process data indicators of test-taking engagement. Presentation 4 provides nuanced insights into the impact of testing session context on test-taking engagement through student qualitative responses. Presentation 5 grows our understanding of how student perceptions of test-taking engagement vary across academic year and major to identify targeted interventions. 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 applications in educational assessment.

Discussant:
Steven Wise (EngagedMeasurement)

Presentations:

1. **Engagement and Item Difficulty in Computer Adaptive Testing: Evidence from Ontario's Grade 9 Mathematics Assessment**
   Hyunah Kim, Education Quality and Accountability Office; Blair Lehman, Brighter Research; Mitch Haslehurst, Nipissing University; Zhimei Gu, Education Quality and Accountability Office; Jennifer Hove, Education Quality and Accountability Office

2. **Fourth Graders' Behavioral Engagement in a Scenario-Based Reading Task: Process Data Insights**
   Burcu Arslan, ETS Research Institute; Yan Fred, ETS Research Institute; Jesse Sparks, ETS; Sarah Rodgers, ETS; Hilary Persky, ETS Research Institute

3. **Beyond Multiple Choice: Understanding Student (Dis-)Engagement in AI-Based Adaptive Dialogue Tasks**
   Marlen Holtmann, IPN - Leibniz Institute for Science and Mathematics Education, EUF – Europa-Universität Flensburg; Marlit Lindner, IPN - Leibniz Institute for Science and Mathematics Education; Libby Gerard, University of California, Berkeley; Marcia Linn, University of California, Berkeley

4. **Student Effort During a Low-Stakes Testing Session: A Mixed Methods Study of Time of Day and Assessment Order Effects**
   Katarina Schaefer, James Madison University; Sara Finney, James Madison University

5. **Peers, Perceptions, and Putting Forth Effort: The Role of Normative Beliefs in Low-Stakes Testing**
   Riley Herr, James Madison University; Dena Pastor, James Madison University; Sara Finney, James Madison University

From Generation to Calibration: Leveraging AI for Item Development, Piloting, and Scoring

Coordinated Poster Session
9:45 AM – 11:15 AM
Intercontinental Los Angeles Downtown, Floor 7: Hollywood Ballroom II

This coordinated session presents evidence on how language models could accelerate the assessment pipeline—from authoring to pretesting to scoring—while safeguarding validity, fairness, and interpretability. Poster 1 introduces a construct-map–guided prompt-engineering workflow for cloning Algebra/Statistics constructed-response (CR) bundles. GPT-5 outputs are human-reviewed, and student responses are used to examine psychometric comparability (e.g., DIF). In Poster 2, LLMs serve as response simulators conditioned on construct-map waypoints and IRT-ability distributions for CR tasks in Visual Reading Comprehension and Statistical Reasoning, examining textual and psychometric comparability. Poster 3 replicates feature-based and fine-tuned-LLM prediction of item difficulty in an early-childhood credential context (~400 MC items), testing transferability and potential reductions in pretest burden. Poster 4 proposes a three-agent scoring framework—annotator, evaluator, decision maker—applied to ~900 middle-school science responses, tracing score changes and improving rubric alignment. Similarly, poster 5 offers a multi-agent system that generates majority-vote scores from five LLMs along with uncertainty. A trial with approximately 300 Grade-4 science responses reached 88% accuracy (Δ20% improvement over single-agent baseline). Collectively, these studies show that, when deployed properly, AI has potential to expand construct-aligned item pools, reduce piloting demands via principled synthetic pretesting, provide preliminary difficulty estimates, and strengthen automated scoring validity.

Presentations:

1. **Constructed-response Item Generation using LLMs in Secondary-school Math Assessments.**
   Hyemin Park, UC Berkeley; Alexis Fernandez, UC Berkeley; Yukie Toyama

2. **Response Generation using LLMs informed by IRT and Construct Maps.**
   Yukie Toyama; Alexander Blum, Stanford University; Hyemin Park, UC Berkeley

3. **Replicating Feature-based and LLM/SLM Fine-tuned Difficulty Prediction in an Early Childhood Education Credential Setting**
   Richard Brown, West Coast Analytics

4. **Increasing Automatic Scoring Validity with a Multi-Agent LLM Framework.**
   Mingfeng Xue, University of North Carolina Greensboro

5. **Enhancing Scoring Accuracy with Multi-Agent Systems: Analyzing Elementary Students' Written Responses in Science Assessment**
   Namsoo Shin; Xunlei Qian, Michigan State University; Harvey Li, Michigan State University; Yucheng Chu, Michigan State University; Cory Miller, Michigan State University; Jiliang Tang, Michigan State University; Krajcik Joseph, Michigan State University

Integrating AI into Assessment and Psychometric Practice

Individual Paper Session
9:45 AM – 11:15 AM
Intercontinental Los Angeles Downtown, Floor 5 : Silver Lake A

Chair:
Jae Jun Jong

Discussant:
Syed Abdul Hadi (UT System and Wisconsin Center for Education Research (WIDA))

Presentations:

1. **AI-Based Scale Development: Identifying Factors in Subjective Topics**
   Eunji Lee, University of Georgia; George Engelhard, University of Georgia; Laura Lu, University of Georgia; Yoojoong Kim, University of Georgia

   This study investigates the application of AI in attitude scale development. Building Attitude Toward Censorship scale, AI generated new items through prompt engineering and organize them into thematic categories. In our dataset, two- or three-factor models demonstrated superior fit compared to the AI-generated five-factor model by EFA and CFA.

2. **Flexible Multimodal Fusion Model: Predicting Item Parameters Using Text, Images and Metadata**
   Hotaka Maeda, Smarter Balanced; Yikai Lu, University of Minnesota Twin Cities

   We built a flexible item parameter prediction model that can utilize all available data, including the text, images, and meta data. AI vision and language models were combined as a deep learning fusion model. Notably, the model handles any item type (e.g., zero to multiple response options, passages, and images).

3. **Beyond Detection: AI-Driven Clustering for Fairness**
   Ismail Dilek

   This study explores the use of AI-driven clustering alongside traditional differential item functioning (DIF) analyses to proactively identify fairness concerns in large-scale assessments. Simulated results demonstrate that AI uncovers latent subgroup patterns beyond conventional methods, offering a proactive equity framework for educational measurement.

4. **Automated test assembly via large language models**
   Sonya Powers, Edmentum; Ege Otenen, Indiana University; Yi He, Edmentum

   We compare the construction of brief quizzes via LLMs, random item selection, optimized item selection, and SME item selection in terms of how well the resulting forms meet the desired content constraints. Results show that LLMs may be challenged when there are multiple constraints to balance simultaneously.

Leveraging Mixed Methods

Individual Paper Session
9:45 AM – 11:15 AM
Intercontinental Los Angeles Downtown, Floor 6: Majestic

Chair:
Margaret de Leon (University of Toronto)

Discussant:
Yasmene Kimble (Stony Brook University)

Presentations:

1. **Generative AI Support in L2 Speaking Assessments: A Mixed-Methods Study**
   Dan Song, The University of Iowa; Alexander Tang, University of Hawaii - Manoa

   This study investigates the role of generative artificial intelligence (GenAI) in second language (L2) speaking assessments. Using a convergent mixed-methods design, we compare student responses with and without ChatGPT assistance. Results highlight AI's affordances and limitations, raising questions about fairness, pedagogy, and ethics in standardized speaking assessment.

2. **Centering Rights-Holders From Co-Design To Identifying Impact in K-12 Assessments**
   Nisha Acharya Julien, Center for Measurement Justice; Sandra Cruz, George Washington University; Maria Hamdani, The Center for Measurement Justice

   Grounded in interview and focus group data, we detail a rights-holder (RH) centered, co-design item development process for 8th grade culturally responsive (CR) math items and RH reactions to these items. Findings revealed positive emotional reactions and increased student engagement. Recommendations support refining CR assessment practices.

3. **LLM-Assisted Coding of Spatial Reasoning Think-Aloud Data: Reliability, Validity and Fairness**
   Jujia Li, University of Alabama; Kaiwen Man, University of Alabama; Wei Huang, Universtiy of Alabama, College of Education; Joni Lakin, University of Alabama

   This study developed a coding framework to transform children's think-aloud protocols during spatial reasoning tasks into structured data. Using large language models , we evaluated reliability, validity, and fairness. By mapping cognitive and metacognitive codes for hypothesized cognitive strategies, the findings bridged qualitative protocol analysis with psychometric modeling.

4. **Keyless Flipped Matrix IRT Reimagines Measurement of Diagnostic Decisions under Uncertainty**
   Ting Wang; Martin Pusic, American Board of Medical Specialty

   Conventional IRT collapses clinical nuance into single key scores. Using a flipped matrix, key less graded response model, 26 pathologists graded 120 prostate biopsies. Discrimination and threshold parameters, visualized via decision probability (PCC) and precision (PIF) curves, revealed clinically meaningful differences undetected by weighted κ. The approach offers scalable assessment to complement expert panels.

5. **Item Logic: Explicitly Explaining How Item Designs Prompt Test Taker Cognition**
   Marjorie Wine, Accessible Teaching, Learning and Assessment Systems (ATLAS), University of Kansas; Alexander Hoffman, AleDev Research & Consulting

   Item logic (the idea for how an item elicits evidence of proficiency) is usually implicit—limiting scalability, training, and item quality. This paper explores how explicit item logic improves item validity, supports professional judgment, and contributes evidence to support the interpretability of test scores for the proposed uses of tests.

Methodological Studies on IRT Metrics for Measuring Growth and Change

Coordinated Paper Session
9:45 AM – 11:15 AM
Intercontinental Los Angeles Downtown, Floor 5 : Silver Lake B

Although item response theory (IRT) metrics are typically constructed from cross-sectional data collection designs (i.e., between students within a fixed grade level) and calibrations, the resulting metrics are frequently used for longitudinal interpretations, such as quantifying student or group-level growth across grade levels. In psychological and health outcomes research, where identical measures are more commonly administered repeatedly over time, it has become widely appreciated that between- and within-person dynamics are often very different, and that measurement representations need to be sensitive to such differences, such as through use of multilevel models (e.g., McArdle, et al. 2014; Reise, et al, 2005; Vogelsmeier et al., 2023). While conditions similar to these are likely also present in educational measurement, such differences are not often an explicit consideration when building vertical scales, perhaps due to the use of designs that (by necessity) entail the administration of different tests across grade levels.

This coordinated session comprises a collection of empirical and simulation studies from investigators at different universities that seek to shed light on the complex dynamics associated with the use of measures typically designed for cross-sectional purposes as a basis for representing growth/change at either individual or group levels.

Discussant:
Sanford Student (University of Delaware)

Presentations:

1. **Item-level Treatment Effect Heterogeneity in Pre-/Post Designs: Comparing Unipolar and Bipolar Metrics**
   Qi Huang, Purdue University

2. **Scaling Multidimensional Tasks for Longitudinal Educational Measurement: Early Reading Development Using ROAR**
   Tongtong Zou, Stanford University

3. **Harnessing Longitudinal Data for Vertical Scale Construction and Calibration**
   Sarah Wellberg, University of Virginia

4. **Cross-Sectional and Longitudinal Item Discrimination in Growth Measurement**
   Xiangyi Liao, University of British Columbia

5. **Evaluating Growth in the Presence of Proficiency-Based Slipping and Guessing**
   Lionel Meng, University of Wisconsin - Madison

Practical issues in CAT

Individual Paper Session
9:45 AM – 11:15 AM
Intercontinental Los Angeles Downtown, Floor 5 : Westwood

Chair:
Ye Ma (Amazon Web Services)

Discussant:
Denis Federiakin (Johannes Gutenberg University of Mainz; Goethe University of Frankfurt)

Presentations:

1. **Improving Polytomous Item Online Calibration via Random Phase Tuning**
   Zhuoran Wang, NCSBN; William Muntean, National Council of State Boards of Nursing; Joe Betts, NCSBN

   A two-phase calibration—random followed by adaptive—is ideal for dichotomous items online calibration, addressing mistargeting and bias. With an ideal-length random phase, items leverage between broad ability coverage and sufficient information. In this way, it is worth exploring the optimal random phase length for polytomous item online calibration.

2. **Integrating Omission Behaviors in Adaptive Testing with IRTree Models**
   Lixin Wu, University of Illinois at Urbana-Champaign; Justin Kern, University of Illinois at Urbana-Champaign; Huan Kuang, Florida State University

   Conventional IRT models often score omissions as incorrect, biasing ability estimates. This simulation compared a 2PL IRT model with two IRTree models (bifactor and multi-unidimensional) in adaptive testing. IRTree models yielded estimates with higher fidelity, reduced bias, and robust RMSE performance, while maintaining acceptable test overlap—ultimately supporting fairer educational assessments.

3. **Exploring Response Time-Based Approaches to Enhance Item Selection in Time-Constrained CATs**
   Joyce Xinle Liu, University of Alberta; Doris Abroampah, University of Alberta; Okan Bulut, University of Alberta

   This post-hoc simulation study utilized large-scale assessment data to evaluate the efficacy of maximum Fisher information for item selection in computerized adaptive tests (CATs) against three response time-based alternatives. Findings indicate that integrating response times enhances ability estimation, test efficiency, and item pool utilization, providing valuable insights for time-constrained CATs.

4. **Implementing Multistage Testing Using Existing Computerized Adaptive Testing System with Minimal Changes**
   Kyung (Chris) Han, GMAC; Sung-Hyuck Lee, GMAC

   Staged CAT (SCAT) enables MST-like navigation within existing CAT systems. Two staged item-block selectors (SCAT1, θ-spread SCAT2) matched CAT/MST accuracy in simulations while broadening within-stage difficulty. For programs, SCAT offers a practical, low-cost path to modernize candidate experience while preserving controls and avoiding costly platform rebuilds.

The Revision of the Testing Standards: Foundations (Invited Joint AERA/NCME Session)

Organized Discussion
9:45 AM – 11:15 AM
Intercontinental Los Angeles Downtown, Floor 5 : Wilshire Grand Ballroom III

This is the first of three joint sessions designed to update the memberships of AERA and NCME and invite input on the revision of the Standards for Educational and Psychological Testing. This session will focus on the Foundations chapters, with particular attention to validity, reliability, and fairness.

The Joint Committee (JC) has been working on revising the 2014 Standards. In this opening session, participants will receive an update on the revision process and learn about major changes proposed for the Foundations chapters. Most importantly, the session will provide time to hear directly from participants—their reactions to the updates as well as their questions, hopes, and concerns about the ongoing revision.

Chair:
Ye Tong (National Board of Medical Examiners)

Presentations:

1. **The Revision of the Testing Standards: Foundations (Invited Joint AERA/NCME Session)**
   Ye Tong, National Board of Medical Examiners; Stephen Sireci, University of Massachusetts Amherst; Pohai Kukea Shultz, University of Hawaii - Manoa; Michael Russell, Boston College; Chad Buckendahl, ACS Ventures, LLC; Andy de los Reyes; Kristen Huff, Curriculum Associates

   This is the first of three joint sessions designed to update the memberships of AERA and NCME and invite input on the revision of the Standards for Educational and Psychological Testing. This session will focus on the Foundations chapters, with particular attention to validity, reliability, and fairness.
   The Joint Committee (JC) has been working on revising the 2014 Standards. In this opening session, participants will receive an update on the revision process and learn about major changes proposed for the Foundations chapters. Most importantly, the session will provide time to hear directly from participants—their reactions to the updates as well as their questions, hopes, and concerns about the ongoing revision.

Individual eBoards: Methodological Investigations in DCM, G Theory, and IRT

Electronic Board Session
11:30 AM – 12:30 PM
Intercontinental Los Angeles Downtown, Floor 5 : Hancock Park

Presentations:

- **Prior Sensitivity in Bayesian Estimation of Model-Based Oral Reading Fluency Scores (eBoard 1)**
  Yusuf Kara, University of Miami; Xin Qiao, University of South Florida; Cornelis Potgieter, Texas Christian University; Joanne Joo, Southern Methodist University

  Model-based oral reading fluency scores are estimated by fitting a binomial-lognormal joint model to passage-level accuracy and speed data. A fully Bayesian estimation approach is efficient; however, sensitivity to priors may be a concern. This study explores the effects of using ability hyperparameter priors with varying levels of information.

- **Bayesian Network Modeling of Dynamic Associations Between Persistence and Learner and Task Characteristics (eBoard 2)**
  Teresa Ober, ETS; Caitlin Tenison, ETS; Diego Zapata-Rivera, ETS; Beata Beigman Kelbanov, ETS

  We introduce a Bayesian network model of student persistence, integrating learner and task factors. Parameterized with real student data from an interactive reading platform, the model could inform adaptive assessment design. The model demonstrates some sensitivity to behavioral and performance patterns reflecting different persistence profiles based on actual students.

- **Cognitive Diagnostic Computerized Adaptive Testing Using Stochastic Curtailment (eBoard 3)**
  Hsiu-Yi Chao, Soochow University; Jyun-Hong Chen, National Cheng Kung University

  To improve CD-CAT efficiency, this study introduces CD-CATSC, integrating stochastic curtailment with PWKL item selection via an innovative probability approach. Results show the method reduces average test length by approximately 22% with a slight decrease in classification accuracy, offering a valuable trade-off between efficiency and precision.

- **Predicting Longitudinal Dyslexia Risk Across Kindergarten Through Grade 3 (eBoard 4)**
  Patrick Kennedy, University of Oregon; Brian Gearin, University of Oregon; Gina Biancarosa, University of Oregon

  Most US states legislate that schools screen students for dyslexia. Early literacy screeners have been shown to be predictive of dyslexia classification over one to two academic years. This study extends prior research to examine the extent to which those measures are predictive longitudinally from kindergarten through Grade 3.

- **A comparison between cognitive diagnosis models and multidimensional IRT using cross validation (eBoard 5)**
  Wenchao Ma, University of Minnesota; Yiming Chen, University of Minnesota; Nana Kim, University of Minnesota

  Cognitive diagnosis models and multidimensional IRT can both provide multidimensional feedback, but the empirical comparison between them is lacking. We compared them across multiple datasets using the cross-validation. By fitting different models, we evaluate their out-of-sample predictive performance (accuracy, F1, AUC, log loss).

- **Item-Level Model Selection for CDMs Using Reversible Jump MCMC (eBoard 6)**
  David Arthur, University of Washington

  Current item-level model selection methods for CDMs either rely on asympotics or do not allow for selection of certain models. We propose a novel reversible-jump MCMC algorithm for performing item-level model selection that addresses these two shortcomings.

- **Confidence Interval Estimation in Multivariate Generalizability Theory: A Bootstrap Approach (eBoard 7)**
  Stella Kim, University of North Carolina Charlotte; Sungyeun Kim, Incheon National University; Qiao Liu, UNC Charlotte

  It is important to understand sampling variability in generalizability theory estimates. However, little research has been conducted, particularly for multivariate generalizabilit theory. The present study examines and compares the performance of several bootstrap estimation methods, building on the work of Tong and Brennan (2007) and Jiang et al. (2022).

- **The Impact of Basal and Ceiling Rules on Reliability Coefficients (eBoard 8)**
  Huan Liu, Riverside Insights; Hyeonjoo Oh, Riverside Insights; Hsin-Ro Wei, Riverside Insights; Min Liang, National Board of Osteopathic Medical Examiners (NBOME)

  This study explores how basal and ceiling rules affect reliability estimates in adaptive testing. Various thresholds, reliability indices, test lengths, sample sizes, and examinee characteristics are systematically examined using simulated item responses. Results clarify the trade-offs between test efficiency and reliability accuracy, to identify optimal configurations in adaptive assessments.

- **Estimating Test Score Reliability in Mixed-Format Assessments with Varying Format Correlations (eBoard 9)**
  Fen Fan, National Board of Medical Examiners; Chunyan Liu, National Board of Medical Examiners

  This simulation study evaluates score reliability in mixed-format tests under varying format correlations. Results show that more polytomous items and higher correlations improve reliability, with multidimensional IRT reliability and stratified alpha outperforming other methods—highlighting their effectiveness in modeling multidimensional structure.

- **Comparing IRT Estimators' Pass/Fail Classification Accuracy: A Simulation Study (eBoard 10)**
  Nicole Bonge, American Board of Internal Medicine; Siyu Wan, ABIM

  This simulation study explores the accuracy of six IRT estimators in classifying examinees as passing or failing across various test lengths, standards (passing scores), and IRT models. Results suggest that the choice of estimator leads to trade-offs in terms of true pass and fail rates, depending on the standard's location.

- **Effect of Modeling Careless Responding with IRTrees on Item and Person Parameters (eBoard 11)**
  Mohammed Abulela, MetaMetrics, Inc. and University of Minnesota; Justin Kern, University of Illinois at Urbana-Champaign

  We used IRTrees to model careless responding in survey data due to its effects on item and person parameters. We utilized a large dataset of the Big Five Personality Traits. Overall, we found that IRTrees resulted in lower standard errors for theta estimates, leading to more precise scoring of individuals.

- **Evaluating Collaborative Filtering for Missing Data Imputation in Large-Scale Assessments (eBoard 12)**
  Sohee Kim, University of South Alabama; Hyunjun Kim, City University of Hong Kong

  Missing data challenges score validity in large-scale assessments. This study compares collaborative filtering (CF) methods with traditional imputations using 2PL IRT simulations across varied conditions. Evaluated via RMSE and MAE, CF demonstrates superior accuracy in large, sparse datasets, offering practical guidance for improving educational measurement.

- **Examining the Mechanisms of the Testlet Effects for OSCE Stations (eBoard 13)**
  Nai-En Tang, National Board of Chiropractic Examiners; Igor Himelfarb; Bowen Wang, National Board of Chiropractic Examiners

  This study demonstrates that the Rasch Random-Effects Testlet Model (RRTM) and the Rasch Fixed-Effects Testlet Model (RFTM) are useful for identifying mechanisms contributing to testlet effects in OSCE stations. Insights gained through this process can guide item refinement, ultimately enhancing the assessment of examinees' clinical competency.

- **Investigation of Concurrent vs. Separate Calibration Methods for a Licensure Practice Assessment (eBoard 14)**
  Authors:Winona Vesey, Kaplan North America; John Denbleyker

  This study proposes comparing separate and concurrent IRT calibrations in a computerized adaptive licensure exam on several different facets. In comparing pass/fail decisions, item analysis, parameters estimates, and algorithm performance across calibration methods, this research aims to evaluate the psychometric implications when using a unidimensional approach to inherently multidimensional data.

- **Practical Guidelines for Estimating Model-based Oral Reading Fluency (eBoard 15)**
  Xin Qiao, University of South Florida; Cornelis Potgieter, Texas Christian University; Akihito Kamata; Sarunya Somsong, Chulalongkorn University; Yusuf Kara, University of Miami; Kuo Wang, Southern Methodist University

  The computer-based oral reading fluency (ORF) assessments are important tools for identifying at-risk students in various educational settings. We conducted a simulation study to explore conditions, such as the number of passages and truncated data, required to estimate model-based ORF scores adequately.

- **Balancing Model Choice and Item Weighting in Mixed-Format Assessment Calibration (eBoard 16)**
  Xiuyuan Zhang, The College Board

  This study empirically compares 2PL and 3PL calibrations of multiple-choice items, combined with graded IRT models for free-response items, under different weighting schemes. Using operational mixed-format data, we examine impacts on calibration accuracy, equating precision, and classification consistency, offering practical guidance on model choice and item weighting in large-scale assessments.

- **Applying Collaborative Filtering for Missing Data Imputation in Large-Scale Educational Assessments (eBoard 17)**
  Sohee Kim, University of South Alabama; Yulim Kang, Yonsei University

  This study examines collaborative filtering methods—SVD and NMF—for imputing missing responses in large-scale assessments (PIRLS). Compared to traditional approaches, CF's performance under varying missingness patterns will be evaluated using RMSE, MAE, classification accuracy, and ability estimates, offering practical guidance for improving validity and fairness in testing.

- **Are We Done Yet? Shortening a Fourth-Grade Math Assessment with IRT (eBoard 18)**
  Jerry Nelluvelil, Harvard Graduate School of Education

  This study investigates whether a fourth-grade math assessment module can be shortened without compromising psychometric quality. Using reliability prophecy, a 2PL IRT model, and cross-grade linking, I identify candidate items for removal. Results show shortened forms maintain reliability and coverage, with implications for fairness, efficiency, and instructional use.

- **Bootstrapping to Determine the Optimal Number of Clusters (eBoard 19)**
  Jason Bryer, City University of New York

  Determining the optimal number of clusters for cluster analysis is challenging. This study explores a new metric that estimates to overlap of cluster centers from bootstrap samples. As compared to six existing metrics, the overlap fit identifies the correct number of clusters in a dataset where cluster membership is known.

- **Detecting Rater Centrality in Human Scores (eBoard 20)**
  Yangmeng Xu, Pearson; Edward Wolfe, Iowa Testing Programs / University of Iowa

  This study explored a cumulative approach for detecting rater centrality in human scores using p values associated with F statistics. We evaluated how quickly (i.e., the number of student responses needed) and how accurately (i.e., Type I and II error rates) it detects centrality among human scores.

Evaluating Assessment Practices for Students with Disabilities (Historically Marginalized Groups SIGIMIE Session)

Coordinated Paper Session
11:30 AM – 12:45 PM
Intercontinental Los Angeles Downtown, Floor 5 : Ladera Heights

This session brings together three data-rich studies that offer fresh perspectives on how students with disabilities (SWDs) interact with large-scale assessments, moving from broad system-level trends to granular behavioral insights and nuanced strand-level analysis. The first paper presents a longitudinal analysis of state-level participation, performance, and accommodation trends from 2010–2023, revealing wide variability across states and grade levels, and highlighting the evolving role of universal tools in computerized testing environments. The second paper leverages process data from over 3,000 SWDs to examine time-use behaviors during math assessments, showing that regulated behaviors—such as strategic reviewing and tool use—are more predictive of performance than extended time alone. The third paper introduces a strand-level approach to analyzing accommodation bundles on Ontario's literacy test, demonstrating that specific combinations of accommodations yield differential effects across reading and writing strands. Together, these studies form a cohesive narrative: understanding and improving assessment equity for SWDs requires attention to system-wide patterns, individual behaviors, and the nuanced interaction between accommodations and content. This session invites educators, researchers, and policymakers to rethink how assessments are designed, supported, and interpreted to better serve diverse learners.

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

Discussant: Martha Thurlow (Retried)

Presentations:

1. **Longitudinal State-Level Analysis of Accommodations, Participation, and Performance for Students with Disabilities**
   Yi-Chen Wu, National Center on Educational Outcomes, University of Minnesota

2. **Time Use Behaviors of Students with Disabilities on Mathematics Test Items**
   Heather Buzick, ACT

3. **Disaggregated Strand-Level Literacy Outcomes from Bundled Accommodations for Secondary Students with Disabilities**
   Pei-Ying Lin, University of Saskatchewan

Aggregating Information for Measurement Over Time: A Comparison of 3 Approaches

Coordinated Paper Session
11:30 AM – 12:45 PM
Intercontinental Los Angeles Downtown, Floor 5 : Westwood

In today's digital learning environments, students engage with personalized instruction and digital assessments on a regular basis, generating rich streams of data that offer opportunities for continuous measurement. However, individual activities may lack sufficient information to yield reliable and valid proficiency estimates for score reporting. This session explores three innovative approaches for aggregating measurement information over time using data from the i-Ready digital assessment and instruction suite. Each paper applies a different modeling framework—Bayesian Item Response Theory (Glicko-IRT), a hybrid XGBoost and deep learning model, and Elo-based skill estimation—to lesson-level responses and interim assessment scores, aiming to dynamically estimate student ability throughout the school year. Models are evaluated using a common framework, with validity evidence drawn from comparisons to subsequent interim assessment scores. The session culminates in a synthesis paper that examines how each modeling approach conceptualizes and quantifies uncertainty, emphasizing the importance of aligning model selection with the specific context, data characteristics, and reporting needs. This discussion provides practical guidance for choosing models that best support valid and defensible interpretations in dynamic measurement scenarios.

Chair:
Erin Banjanovic

Discussant:
John Behrens (University of Notre Dame)

Presentations:

1. **Glicko-IRT: An Application of Dynamic Measurement**
   Ted Daisher; Erin Banjanovic; Logan Rome, Curriculum Associates

2. **A XGBoost and Deep Learning Hybrid Model for Predicting Student Proficiency**
   Keith Holliday, Curriculum Associates

3. **Elo-Based Skill Estimation for Predicting Reading Growth**
   Michael Asher, Carnegie Mellon University; Ken Koedinger, Carnegie Mellon University; Paulo Carvalho

4. **Context, Accuracy, and Uncertainty: Guidance for Selecting a Dynamic Measurement Model**
   Erin Banjanovic

CANCELLED! When the Data Stops – Implications of Halting Large-Scale Education Data Collections (Committee on Informing Assessment Policy Session)
Organized Discussion
11:30 AM – 12:45 PM
Intercontinental Los Angeles Downtown, Floor 7: Roosevelt B

Chair:
Heather Klesch (Pearson)

Discussant:
Trent Workman (Pearson)

Presenter(s):
Heather Klesch, Pearson; Donna Roper; Trent Workman, Pearson; Andrew Middlestead, Michigan Department of Education; Christopher Domaleski, Center for Assessment; Peggy Carr, Former Commissioner, National Center for Education Statistics, U.S. Department of Education

This panel session will explore the consequences of halting large-scale federal education data initiatives, including the National Assessment of Educational Progress (NAEP) and the Common Core of Data (CCD). Panelists will examine the implications of losing NAEP's 50-year trend data, the disruption to equity monitoring and funding formulas, and the broader impact on research, policy, and accountability. The session will also address the lack of centralized information about recent federal policy changes affecting data access and transparency. Experts from research, policy, and practice will offer insights and policy advice for NCME leadership and practitioners. Attendees will gain practical strategies for navigating data gaps, preserving historical datasets, and leveraging alternative sources. The session aims to equip the measurement community with tools to advocate for data continuity and maintain valid, equitable assessment practices in an era of uncertainty.

Embedding Coherence in Assessment Design and Standard Setting Implementation

Coordinated Paper Session
11:30 AM – 12:45 PM
Intercontinental Los Angeles Downtown, Floor 5 : Silver Lake A

Range Performance Level Descriptors (RPLDs) are gaining momentum as the interpretive foundation for large-scale assessments, providing the cognitive and content-based rationale for score meaning (Perie, 2008; Egan et al., 2012; Forte, 2017, Huff et al., 2025). PLDs define what it means to be At or Above Proficient. As Cizek and Bunch (2007) observed, “standards are set more by the panels who craft the PLDs than by those who rate items or performances” (p. 193). This session explores ways the Embedded Standard Setting (ESS, Lewis & Cook, 2020) process and the Embedded ID-Matching Method (Schneider & Lewis, 2021), which uses the ESS algorithm to avoid transitioning panelists from one cognitive task to another, have been recently operationalized. We overview why and how these extensions are solving educational measurement problems in operational testing programs. Each paper focuses on supporting a validity argument using evidence.

Chair:
Chad Buckendahl (ACS Ventures, LLC)

Discussant:
Chad Buckendahl (ACS Ventures, LLC)

Presentations:

1. **A Framework for Implementing PAD through Embedded Alignment and Standard Setting**
   Ellen Forte, edCount, LLC

2. **Hybrid Standard Setting Design: A Case Study in Implementing ESS for Reading**
   Jessalyn Smith, DRC; Joseph Fitzpatrick, DRC; Deema Abu Abdo, South Carolina Department of Education

3. **Evaluating Validity Evidence from Embedded ID Matching**
   Christina Schneider, Cambium Assessment; Sangdon Lim, Cambium Assessment

4. **Evaluating RP Business Rules on the Cut Score Recovery from Range PLDs**
   Sangdon Lim, Cambium Assessment; Christina Schneider, Cambium Assessment, Inc.

Measurement-Informed Approaches to Evaluate GenAI Outputs (Artificial Intelligence in Measurement and Education SIGIMIE Session)

Organized Discussion
11:30 AM – 12:45 PM
Intercontinental Los Angeles Downtown, Floor 5 : Wilshire Grand Ballroom III

Chair:
John Whitmer

Discussant:
Magdalen Beiting-Parrish

Presenter(s):
Learning Data Insights; Kristen Dicerbo, Khan Academy; Britte Cheng, Menlo Education Research; Nancy Otero, Gates Foundation; Susan Lottridge, Pearson

Generative artificial intelligence has rapidly transformed educational assessment, with increasing adoption across formative and summative evaluation contexts. However, the field lacks consensus on how to apply measurement principles to evaluate GenAI-generated content and scoring. While traditional psychometric concepts of validity, reliability, and fairness remain foundational, their application to GenAI requires the use of new approaches that combine conventional psychometrics with techniques from computer science.

This organized discussion examines how measurement science principles apply to GenAI assessment applications through diverse stakeholder perspectives. The panel explores current practices across academic research, commercial testing, private philanthropy, applied research and development, identifying both promising approaches and persistent challenges in applying psychometric standards to AI-generated content. Key questions include: What validity evidence should be created for different applications of AI? What type of fairness analysis should be conducted to ensure suitability of outputs across diverse student populations? What practical evaluation frameworks can guide assessment developers and educational leaders in implementing GenAI solutions while maintaining measurement quality?

The discussion aims to establish measurement-based evaluation standards and provide actionable guidance for stakeholders developing, procuring, or implementing GenAI assessment solutions. This is the coordinated session of the Artificial Intelligence of Measurement in Education (AIME) subcommittee.

Post Launch Measurement: Monitoring Test and Test-Taking Stability

Coordinated Paper Session
11:30 AM – 12:45 PM
Intercontinental Los Angeles Downtown, Floor 5 : K-Town

Operational testing programs often rely on the assumption that test scores retain consistent properties over time. This stability enables standard psychometric practices such as item anchoring, differential item functioning (DIF) analysis, item parameter drift detection, and equating/scaling procedures. However, advances in artificial intelligence (AI), machine learning (ML) methods, and computing power now enable rapid innovation in item generation, automated scoring, and adaptive delivery. These developments also introduce challenges to preserving score meaning, especially as assessments evolve in structure and content. As a result, modern assessment programs face a central tension between innovation and interpretive continuity. This coordinated session presents methodological and applied research on detecting and responding to changes in score meaning and item behavior following operational test updates. Presentations address a range of threats to stability, including population shifts, AI-generated task types, improved scoring algorithms, and expanded item pools. Collectively, the papers included in this session demonstrate how quasi-experimental and psychometric methods, such as difference-in-differences (DiD) analyses, score stability analyses, item parameter monitoring, test taker behavior analytics, and modern item anchoring strategies, can support continuous test improvement without compromising the properties of test scores.

Discussant:
Chun Wang (University of Washington)

Presentations:

1. **Estimating Causal Score Impacts of Test Updates in Continuous High-Stakes Assessments**
   Manqian Liao, Duolingo; J.R. Lockwood, Duolingo

2. **Evaluating Impact of Test Updates on Score Distributions With an Information Adjustment Approach**
   Siyuan Chen, Duolingo, Inc.; J.R. Lockwood, Duolingo

3. **Maintaining Scale in Continuous Calibration with Item Parameter Uncertainty**
   Steven Nydick, Duolingo; J.R. Lockwood, Duolingo; Manqian Liao, Duolingo; Siyuan Chen, Duolingo, Inc.

4. **Monitoring Shifts in Test Taker Behaviors in a Visual Dashboard**
   Xiaowan Zhang, Duolingo; Yena Park, Duolingo

Redefining Readiness to Recognize Schools' Civic Missions

Coordinated Paper Session
11:30 AM – 12:45 PM
Intercontinental Los Angeles Downtown, Floor 6: Majestic

This session addresses the growing need for assessment systems that provide evidence of graduates' readiness to engage effectively in civic life. The need for expanded civic learning opportunities is evident from recent assessment and survey data, but educators and policymakers have limited access to assessment tools and practices to monitor that learning. As schools aim to prepare young people to thrive in a rapidly changing social, technological, and political landscape, definitions of postsecondary readiness will need to reflect schools' civic missions. The session brings together four complementary papers that collectively map the civic assessment landscape, propose strategies for integrating civics into large-scale systems, examine tensions between policy demands and local practice, and demonstrate learner-focused approaches that capture variation in students' civic development. Presentations will highlight innovations such as performance assessments and person-centered analytic methods, while also exploring challenges of feasibility, political context, and equity of access. A discussant will synthesize these themes and lead an interactive conversation with presenters and participants. By showcasing both system-level strategies and learner-centered perspectives, this session underscores how assessment can support monitoring of civic readiness, guide decision-making, and ensure that public education fulfills its civic mission alongside its academic and economic goals.

Discussant:
Joseph Kahne (University of California, Riverside)

Presentations:

1. **Assessing Civic Learning Opportunities and Outcomes: A Landscape Analysis**
   Samuel Rikoon, American Institutes for Research; Laura Hamilton, National Center for the Improvement of Educational Assessment, Inc.; David Kidd, Harvard University

2. **Designing Large-Scale Assessments to Prioritize Civic Learning Opportunities and Outcomes**
   Christopher Brandt, Center for Assessment

3. **Building Stronger Civic Readiness Measurement Amid Policy and Practice Tensions**
   Maggie Reeves, Urban Institute; Rachel Lamb, Urban Institute

4. **Assessing Proximal Outcomes with Profiles in Civic Learning**
   Mary Cochron, Metimur Educational Measurement

Systematically Advancing the Measurement of Inclusion in Schools Worldwide

Coordinated Paper Session
11:30 AM – 12:45 PM
Intercontinental Los Angeles Downtown, Floor 5 : Silver Lake B

It is currently impossible to compare and align terminology, definitions, and the corresponding measurement of inclusive practices in education contexts globally. The complexity of inclusion at the intersection of culture and contextualized historical precedence has resulted in many wildly diverse measures intended to promote inclusive practices in education. The body of work prepared for this symposium is a first large-scale effort to address this gap in educational measurement of inclusion through a systematic scoping, review, and measurement discourse analysis. We identified 201 distinct measures nested in 235 studies for analysis. We organized measures into 6 categories (Attitudes and Beliefs, Teacher Competency and Practices, Student Perceptions and Experiences. Parental Perceptions and Attitudes, School Climate and Culture. Inclusion Implementation and Assessment, Instructional Strategies and Differentiation, Leadership and Policy). Paper 1 presents findings from a pre-registered systematic review of measures of inclusive practices used in educational contexts worldwide. Paper 2 discusses the psychometric properties of the identified measures, reporting on their reliability, validity, structure, and characteristics. Paper 3 uses AI-assisted discourse analysis to examine how inclusion is framed in both the measurement items and how researchers report on students and findings in their studies. Collectively this work advances the measurement of inclusive practices.

Discussant:
Stephen Sireci (University of Massachusetts Amherst)

Presentations:

1. **A Scoping Review of the Measurement of Inclusive Practices in Schools Worldwide**
   Melissa Stoffers, University of Nevada; Michael McCarthy, Education Collaboratory at Yale University; Christina Cipriano, Yale University

2. **The Properties of Inclusive Measures Used in School Settings Worldwide**
   Sophie Barnes, Yale University; Melissa Stoffers, University of Nevada; Michael McCarthy, Education Collaboratory at Yale University; Christina Cipriano, Yale University

3. **Using AI to Analyze Language in Inclusive Practices Measures**
   Michael McCarthy, Education Collaboratory at Yale University; Sophie Barnes, Yale University; Melissa Stoffers, University of Nevada; Christina Cipriano, Yale University

TACtics: The Future of Technical Advisory Committees for State Testing Programs

Coordinated Paper Session
11:30 AM – 12:45 PM
Intercontinental Los Angeles Downtown, Floor 7: Roosevelt A

State Technical Advisory Committees (State TACs) play an important and often unsung role in the validation and monitoring of state testing programs in the United States. As testing standards evolve and accountability policies shift, TACs, their vendors, and state agencies must remain vibrant and responsive. This symposium presents selected perspectives about the future of State TACs from members serving key roles in State TAC meetings: the TAC member, the state assessment director, the vendor, and the facilitator.

Andrew Ho will open with "the State of the State TAC," including results of his NCME Presidential Initiative to honor over 100 NCME members who serve State TACs. Derek Briggs will argue for higher standards in TAC meetings toward improving transparency, engagement, and evidence. Hope Worsham will offer the state agency's perspective, explaining how she uses TAC meetings to negotiate tensions between psychometric rigor and policy considerations. Jennifer Dunn will present the vendor's perspective, including the challenges and opportunities vendors face when integrating TAC recommendations into test design and analysis. And Juan D'Brot will discuss these presentations from his perspective as a longtime facilitator of State TACs. Presentations will be sufficiently brief to allow at least 30 minutes for engagement with the audience.

Chair:
Susan Lyons (Lyons Assessment Consulting)

Discussant:
Juan D'Brot (Center for Assessment)

1. **The State of the State TAC: Recommendations from NCME's Effort to Honor TAC Service**
   Andrew Ho, Harvard University; Scott Marion, Center for Assessment; Erika Landl, Center for Assessment; Derek Briggs, University of Colorado - Boulder

2. **Meeting the Moment: Seven Recommendations for the Next Generation of State TACs**
   Derek Briggs, University of Colorado - Boulder

3. **TACs from the State Agency Perspective: Cutting the Gordian Knot**
   Hope Worsham, Arkansas Department of Education

4. **State TACs from the Vendor Side: Candor, Constraints, and Craft**
   Jennifer Dunn, The College Board

Contemporary Issues in Scaling, Linking & Equating SIGIMIE

Meeting
12:30 PM – 1:30 PM
Intercontinental Los Angeles Downtown, Floor 6: Royal

The Contemporary Issues in Scaling, Linking, & Equating (SLE) SIGIMIE is dedicated to:

- Exploring emerging measurement challenges in establishing and maintaining score scales
- Upholding fundamental principles by ensuring that our rapidly evolving fields remain grounded in rigorous, principled approaches that are explainable, interpretable, and defensible
- Fostering the integration of theory and practice; and bringing together professionals from diverse backgrounds to exchange knowledge and to develop solutions

Student & Early Professionals Headshots
Social
1:00 PM – 2:30 PM
Intercontinental Los Angeles Downtown, Floor 5 : Hancock Foyer

Thank you to our photo session sponsors:

- ACS Ventures, LLC
- edCount LLC

Individual eBoard: Practical Issues in Assessment Development and Measurement
Electronic Board Session
1:45 PM – 2:45 PM
Intercontinental Los Angeles Downtown, Floor 5 : Hancock Park

Presentations:

1. **Integrating LLM-derived measures with text-based features to predict item difficulty**
   Xuechun Zhou, Ascend Learning; Lucia Liu, Ascend Learning; Hanwook Yoo, Ascend Learning

   Item pretesting is time-consuming and test assembly requires items targeted to a desired ability range. This study investigates whether combining large language model (LLM)-derived measures with text features, using logistic regression with cross validation (LG-CV), can improve predictive difficulty classifications that help identify pretesting items that are most needed for test assembly.

2. **Enhancing Test Development: A Machine Learning Approach to Predicting Science Item Quality**
   Ahmed Bediwy, The University of Iowa; Melinda Taylor

   State assessments need a steady supply of new items to maintain test validity. This study streamlines predicting middle school science item difficulty and quality by combining stored item metadata with LLM-extracted item features across four approaches. Outcomes may vary depending on the purpose and the machine learning model applied.

3. **Examining Boolean Approaches in the Coding of Computer-Administered Item Design Features**
   Eunjung Myoung, Stanford University; Jialu Zhao, Stanford University; Maria Ruiz-Primo, Stanford University; Guillermo Solano-Flores, Stanford University

   We address challenges of coding endeavors with heterogeneous coding objects and many dichotomous variables. We report on a system for coding test item design features, using items from PISA, NAEP, and SBAC. Analyses include inter-coder agreement, association, and Boolean approaches, treating consistency as a conjunction function or equivalence function.

4. **MC Items Derived from CR Item Responses: Better Measurement?**
   Anthony Fina, University of Iowa; Ahmed Bediwy, The University of Iowa

   The measurement properties of constructed response (CR) items and equivalent multiple choice (MC) items are compared. Examples are used to make connections between each item's content and construct representation as it relates to the aligned standard. Implications and practical recommendations for test developers are provided.

5. **Evaluating the Psychometric Robustness of Shortened Certification Exam**
Matthew Scaruto, Ascend Learning; Candice Davis, Ascend Learning; Hanwook Yoo, Ascend Learning

Previous research suggests that reducing exam length can maintain psychometric integrity while enhancing operational efficiency and alleviating stress and cognitive fatigue. This study examines the development and evaluation of shortened test forms of a healthcare certification exam, guided by best practices in test design and psychometric analysis.

6. **Developing Parallel Test Forms with Equivalent Construct Representation**
ci zhang; Xiangdong Yang, East China Normal University

Current parallel test development emphasizes reliability over validity. This study proposes a novel method for generating parallel test forms with equivalent construct representation. An application to mental rotation task confirmed the feasibility of the new method. The study expands the breadth of parallel tests, ensuring both reliability and validity.

7. **Examination of Factors Affecting Students' Reading Ability in PIRLS with Structural Equation Modelling: A Cross-Country Comparison**
ZHEN WANG, Hong Kong Examinations and Assessment Authority

This study uses structural equation modelling to examine the relationships between various factors affecting students' reading comprehension in PIRLS. We investigated the key variables and evaluated the differences of the model parameter estimates across three countries and several sub-groups. The results help further improve the reading environment and school policy.

8. **Two for the Price of One: Large-Scale Assessment Items in State Tests**
Paul Bailey, American Institutes For Research; Martin Hooper, American Institutes for Research; Young Yee Kim, American Institutes for Research

We investigate viability of embedding items from international large-scale assessments into state summative tests to increase efficiency and minimize student burden. To do this we investigate the psychometric properties of administering a reduced number of items to a census of students along with a state assessment.

9. **Exploring Developmental Patterns of General Interest and Career Interest in Computer Science**
Feng Wang; Rongxiu Wu, Harvard University; Susan Sunbury, Harvard University; Philip Sadler, Harvard University; Gerhard Sonnert, Harvard University

Using latent class analysis of U.S. undergraduates, we identified four patterns of computing interest and career intentions: high (CS interest at high school entry and exit) →high (CS career interest), low→low, high→low, and mid→low. Results reveal critical leakage points where sustained interest does not convert into career aspirations.

10. **Measuring problem solving for micro-credentials: Evidence from the Philippines Alternative Learning System**
Farhan Azim, The University of Melbourne; Bruce Beswick, The University of Melbourne

This study validated an assessment instrument measuring the Problem Solving capability within the Philippine Alternative Learning System's Life Skills curriculum. Psychometric evaluation across multiple rater groups established reliability and validity in identifying empirically derived performance levels mapped to observable skill standards, enabling micro-credentialing of this critical employability proficiency.

11. **A Confusion Matrix Approach to Rater Accuracy**
Russell Almond, Florida State University; Andrew Krumm; Blaire Carpenter, Florida State University; Kayla Marcotte, University of Michigan

This paper introduces a model for rater accuracy based on the confusion matrix based through defining the probability of a rater over-or-under-rating a performance. It demonstrates how the confusion matrix can be estimated, and shows it relationship to other measures of rater agreement and severity.

12. **Predicting item response time using natural language processing to reduce speededness**
Aijun Wang; Yu Zhang, The Federation of State Boards of Physical Therapy

This research explores using NLP to predict response times for scenario-based items. By identifying time-intensive items during test development, the study tries to reduce the unintended effects of speededness and improve the validity of test scores.

13. **Evidence from Learning Data and Automated Item Review: Toward AI- and LLM-Assisted Item Evaluation**
Jungwon Kyung, University of Massachusetts, Amherst; Steven Moore, George Mason University

This study integrates learning analytics and automated item review to diagnose student struggles in a university data science course. Using knowledge component modeling, clustering, and rubric-based item flaw detection, the analysis identifies persistent conceptual difficulties and flawed items, generating actionable insights for data-driven instructional and assessment improvement.

14. **Empathy or Attention? A Data-informed Approach to Identify Misresponse to Reverse-coded Items**
Wendy Christensen, University of Colorado School of Medicine; Rachael Tan, University of Colorado School of Medicine; Tai Lockspeiser, University of Colorado School of Medicine

We used the Toronto Empathy Questionnaire to develop data-informed flags to screen for item misresponse to reverse-coded items. By analyzing response patterns from over 500 medical students, we demonstrate how scalable methods to identify item misresponses that are useful in an authentic context can be developed.

15. **Attempts to clean up the messy middle problem: Next Steps**
Zechu Feng, The University of Hong Kong; Jimmy de la Torre, The University of Hong Kong

Messy middle (MM) can affect the validity of inferences from learning progression-based assessments. It has been shown that the MM can result from fitting simpler IRT model to data generated using a more complex IRT model. In this study, another model for generating more complex data is proposed and explored.

16. **Evaluating Sample Size and Demographic Representativeness in Intelligence Norm Development**
Wenjing Guo, Pearson; Troy Courville, Pearson; Lei Shi, Pearson

This study examines how sample size and demographic mismatches affect ability estimates in intelligence assessments. Using simulated and empirical data, findings show larger samples yield more accurate mean and SD estimates, while deviations in education-level distributions distort ability scores, especially at distribution tails. Results inform practices for normative sample design.

17. **Investigating the Effects of Text-to-Speech in Student Achievement**
Daeryong Seo; Eric Moyer, Pearson

This study investigated the effects of text-to-speech on elementary social studies assessments using propensity score matching. Item- and test-level analyses revealed significant score improvements, particularly for lower-achieving students. Findings indicate that text-to-speech reduces reading-related barriers, enhances construct validity, and supports more equitable assessment of social studies knowledge and skills.

18. The Impact of Testing Order on Student Performance and Testing Behavior
kyunghee suh, Cambium Assessment; Ian Campbell, Cambium Assessment

This study investigated how large-scale assessment component order (CAT/PT) affected student performance and testing behavior. Using propensity score weighting on observational data, it found minimal impact of test order on performance or behavior. These empirical findings directly informed and refined existing administration guidelines, ensuring they were empirically supported.

19. Investigating Omitted Response Patterns
Jin Koo, Enrollment Management Association

This study examines omitted response patterns based on the passage category in Reading, the item category in Mathematics, and different characteristics of examinees within both assessments. My analysis showed that the pattern of omit rates varied by item category and characteristics of examinees, while controlling for students' ability.

20. Automating Consequential Validity Ratio: An R Shiny App for Examining Test Fairness
Kushmakar Baral, University of Denver; Yixiao Dong, University of California Santa Barbara

This study introduces an R Shiny app that automates calculations of the Consequential Validity Ratio. Using educational and psychological datasets, we demonstrate the accuracy, efficiency, and practical value of the app. The tool enhances the accessibility of fairness evaluation, supporting the broader integration of consequential validity in test validation practices.

21. Analyzing Construct Representation with Case Lists
Dylan Ward, American Board of Obstetrics and Gynecology; Anthony Moreno-Sparks, American Board of Obstetrics and Gynecology; Pooja Shivraj, American Board of OBGYN; Heath Kincaid, American Board of Obstetrics and Gynecology

The American Board of Obstetrics and Gynecology utilizes candidate-submitted, examiner-selected case lists to supplement structured cases during its certification exam. As such, we developed a novel method to investigate how well their combination managed to represent the full construct of OB-GYN.

22. The Impact of Disruptions on Test Scores
Shonai Someshwar, National Council of State Boards of Nursing; William Muntean, National Council of State Boards of Nursing; Joe Betts, NCSBN

This study examined the impact of internet-related disruptions on candidate performance during a high-stakes computer adaptive licensure exam. Using stratified sampling, groups of disrupted and non-disrupted candidates were compared within and across restart conditions. T-tests, ANOVA, and hierarchical change-point regression models revealed no statistically or practically significant differences in scores.

23. Impact of Essay Length on Human Scoring and Chain-of-Thought-Based Automated Scoring Systems
Ayfer Sayin, Gazi University; Okan Bulut, University of Alberta

This study examined how essay length influences human scoring and automated scoring with chain-of-thought prompting. Analyzing Turkish essays, results revealed that longer essays tend to receive higher scores. These findings highlight the importance of establishing optimal word ranges and investigating whether higher scores reflect genuine content richness or bias.

24. **Evaluation Criteria for Measuring the Accuracy of Generative AI Chatbots for Teacher Learning**
    Jamie Mikeska, ETS; Beata Beigman Klebanov; Shreyashi Halder, ETS; Heather Jorgenson, George Washington University; Kashish Behl, ETS; Aakanksha Bhatia, ExcelOne

    In this study, a cross-disciplinary research team examined the application and use of six evaluation criteria for measuring the accuracy of GenAI digital teaching simulations for use in supporting educator learning of key teaching competencies.

Advances in Scoring English Language Learner Items and Responses

Coordinated Paper Session
1:45 PM – 3:00 PM
Intercontinental Los Angeles Downtown, Floor 7: Roosevelt A

This session presents four studies on advancing automated scoring for educational assessments for English Language Learners, focusing on both verbal and written responses. For verbal scoring and crisis response detection, the research explores moving beyond transcription-based approaches to ones that include direct modeling using the student audio as input. This approach has the benefit of including prosodic elements into the scoring itself beyond just the content of the response. For written assessments, the studies investigate optimal model training strategies which include prompt-specific or prompt-agnostic modeling and data augmentation. Collectively, these papers highlight key innovations in making automated scoring more accurate, robust, and safe by leveraging richer data sources and more sophisticated modeling techniques.

Discussant:
Arthur Thacker (Human Resources Research Organization)

Presentations:

1. **Automated Scoring of Verbal Responses in ELPA21: Comparing Models**
   Christopher Ormerod, Cambium Assessment

2. **Detecting At-Risk Students from Verbal Responses**
   Gitit Kehat, Cambium Assessment

3. **Comparing Prompt-Agnostic and Prompt-Specific Automated Scoring Models for WIDA Writing Prompts**
   Gregory Jacobs, Pearson

4. **Enhancing Automated Scoring Model Performance through Data Augmentation: Evidence from WIDA Writing Assessments**
   Justin Barber, Pearson

Automated Coding with LLM: Accuracy and Fairness

Coordinated Paper Session
1:45 PM – 3:00 PM
Intercontinental Los Angeles Downtown, Floor 5 : Silver Lake B

The rapid advancement of large language models (LLMs) has generated growing interest in their potential to support adaptive, conversational assessments targeting 21st-century skills such as collaboration and communication. A critical prerequisite for extracting meaningful evidence from conversation data is the ability to code communication into predefined categories of corresponding frameworks. LLMs offer promising opportunities to automate this labor-intensive coding process diverse assessment contexts. In this coordinated session, we present findings from four recent studies that explore the use of ChatGPT for automated coding of communication and interview data. We examine the accuracy of ChatGPT coding compared to human raters, evaluate their consistency across tasks, and discuss design considerations such as prompt formulation and contextual adaptation. Importantly, we also address the fairness of ChatGPT-based coding, analyzing whether performance varies across demographic groups. Our session highlights both the potential and limitations of LLMs in improving assessment efficiency and scalability, while emphasizing the need for rigorous validation to ensure responsible and equitable use.

Discussant:
Victoria Yaneva (National Board of Medical Examiners)

Presentations:

1. **Reliability and Fairness of LLM-based coding on communication data**
   Jiangang Hao, ETS; Wenju Cui, ETS; Patrick Kyllonen, ETS; Emily Kerzabi, ETS

2. **ChatGPT-based Automated Coding for Hierarchical Coding Framework**
   Wenju Cui, ETS; Jiangang Hao, ETS; Patrick Kyllonen, ETS; Emily Kerzabi, ETS

3. **An exploration of human-AI coding across different types of qualitative data**
   Yuan Wang, ETS; Wenju Cui, ETS; Reginald Gooch, ETS; Jiangang Hao, ETS; Guangming Ling, ETS; Sara Haviland, ETS

4. **Leveraging Large Language Models for Automated Coding of Socially Shared Regulation of Learning**
   Yang Jiang, ETS; Yi Song, ETS Research Institute; Chunyi Ruan, ETS

Combining Through-Year Scores: Operational Approaches, Challenges, and Practical Implications

Coordinated Paper Session
1:45 PM – 3:00 PM
Intercontinental Los Angeles Downtown, Floor 5 : Ladera Heights

A number of states are considering the affordances and weighing the challenges of implementing a through-year assessment system. Through-year programs assess students at multiple points during the year, providing more frequent information while also generating a single summative score for state accountability purposes. There is no clear consensus, however, on which approaches to summative score generation best support different score interpretations that states may wish to make. In fact, there is very little research in this area.

Discussant:
Nathan Dadey (The National Center for the Improvement of Educational Assessment)

Presentations:

1. **A Research Synthesis and Agenda for Score Aggregation for Through-Year Assessment Programs**
   Nathan Dadey, The National Center for the Improvement of Educational Assessment

2. **Creating Summative Scores from Instructionally Embedded Results**
   Jake Thompson, ATLAS, University of Kansas; Brooke Nash, University of Kansas

3. **Summarizing Achievement with Linear Weights: What Works—and What Can Go Wrong**
   Garron Gianopulos, Cambium Assessment

4. **Evaluating Scoring Options for Montana's Aligned to Standards Through-Year Mathematics Assessment**
   Kyla McClure, University of Colorado Boulder

Empirically Evaluating Claims of Instructional Usefulness

Coordinated Paper Session
1:45 PM – 3:00 PM
Intercontinental Los Angeles Downtown, Floor 5 : Westwood

Evans and Marion (2024) defined and conceptualized instructionally useful assessments and defined such assessments as those that "…provide substantive insights about student learning strengths and needs relative to specific learning targets that can positively influence the interactions among the teacher, student, and the content" (p. 19). They identified ten features that could explain why some assessments, and the information they produce, are more or less likely to inform instruction productively. Marion and Evans (2025) later outlined several approaches for empirically evaluating the instructional usefulness of assessments.

Following a framing paper, this symposium comprises three empirical studies that examine the instructional usefulness of two different types of assessments. The first two papers report on a pilot study examining the inferences and intended actions by a set of educators using Curriculum Associates' i-Ready assessment suite. The third paper reports on the perceived instructional usefulness of a formative writing assessment used in Gwinnett County Public Schools.

Discussant:
Derek Briggs (University of Colorado - Boulder) and Kristen Huff (Curriculum Associates)

Presentations:

1. **A Framework for Evaluating Instructional Usefulness**
   Scott Marion, Center for Assessment; Carla Evans, Center for Assessment

2. **Exploring the Interpretations and Uses of Interim Assessment Score Reports**
   Nicolas Buchbinder, University of Colorado Boulder

3. **An Interpretation and Use Argument to Support an Evaluation of Instructional Usefulness**
   Daniel Raphael, Boston College

4. **Validating the Instructional Usefulness of Formative Writing Assessments**
   Elizabeth Blackmon, Gwinnett County Public Schools

Featured Session: Threats and Opportunities on the Federal Statistics Landscape

Invited Session
1:45 PM – 3:00 PM
Intercontinental Los Angeles Downtown, Floor 5 : Wilshire Grand Ballroom III

The statistical system in the United States stands at a critical crossroads. Changes in governance, funding, and leadership at the federal statistical agencies, including the National Center for Education Statistics, present both serious threats and significant opportunities. In this panel, leaders from the American Statistical Association (ASA) and the Council of Professional Associations for Federal Statistics (COPAFS) and a former Commissioner of the National Center for Education Statistics (NCES) join in a discussion that highlights the evolving landscape for federal statistics in general and education statistics in particular. The role and requirements for federal statistical agencies in a well function democracy are explored and discussed in relation to current events. ASA's initiatives on the Nation's Data at Risk and the re-imagining of NCES are highlighted, and opportunities for engagement and action are discussed.

Chair:
Rich Patz

Panelists:
Peggy Carr, Michelle Crosby, Paul Schroeder

Measurement Challenges in Assessing SEL Competencies in International Longitudinal Studies

Coordinated Paper Session
1:45 PM – 3:00 PM
Intercontinental Los Angeles Downtown, Floor 6: Majestic

The purpose of this coordinated paper session is to discuss some of the unique measurement challenges associated with large-scale international K-12 assessments of SEL skills. Studies included in the session administered numerous SEL measures along with achievement and other measures to large longitudinal or repeated cross-sectional samples. Each study identifies measurement challenges related to this year's NCME themes (e.g., tools for measuring/ improving fairness; interdisciplinary innovations; psychometric pain points) and illustrates approaches to address these challenges. Studies cover K-12 and early childhood, discuss cross-sectional and longitudinal designs, skills trajectories, different constructs, statistical response style adjustments, forced-choice versus rating-scale methods, and correlates of change. A theme across the session will be a sharing across presenters and discussant of lessons learned and psychometric methods that might help to increase measurement quality and proper interpretation of findings from large-scale international K-12 assessments of SEL skills.

Discussant:
Daniel Bolt (University of Wisconsin - Madison)

Presentations:

1. **Unraveling Developmental Social-Emotional Skills Trajectories: Evidence from the SENNA Longitudinal Study**
   Ricardo Primi, Universidade São Francisco; Ana Crispim, Ayrton Senna Institute; Oliver John, University of California, Berkeley; Filip De Fruyt, Universiteit Gent; Daniel Santos, University of São Paulo, Ribeirão Preto; Luiz Scorzafave, University of São Paulo, Ribeirão Preto; Ana Zuannazi, Ayrton Senna Institute; Karen Teixeira, EduLab21, Ayrton Senna Institute; Gisele Alves, Edulab21, Ayrton Senna Institute, Brazil

2. **Social-Emotional and Skills Growth of Young Children: Thailand's Newborn Cohort Study**
   Weerachart Kilenthong, University of the Thai Chamber of Commerce (UTCC); Sartja Duangchaiyoosook, University of the Thai Chamber of Commerce (UTCC)

3. **Ratings versus Forced-Choice SEL Skill Measurement: Skills Trajectories Implications**
   Patrick Kyllonen, ETS; Stephanie Carlson, University of Minnesota; Thomas Dohmen, University of Bonn; James Heckman, University of Chicago

Multi-Modal Approaches to Test Security: Patterns, Similarities, and Biometrics (Test Security SIGIME Session)

Coordinated Paper Session
1:45 PM – 3:00 PM
Intercontinental Los Angeles Downtown, Floor 5 : Boyle Heights

This coordinated session presents four innovative approaches to detecting misconduct in digital assessment environments. Belzak examines how “jagged” subscore profiles across language skills correlate with validity flags in remote proctoring contexts, finding that discrepancies spanning two CEFR proficiency bands significantly increase misconduct probability. Grabovsky and Feinberg investigate Random Forest classification methods for detecting collusion clusters, exploring threshold selection techniques that balance sensitivity and specificity while accounting for operational constraints and expected cheating prevalence. Eckerly and Gorney introduce computational efficiencies for answer similarity analysis using the M4 statistic through lookup table approaches, making comprehensive pairwise examinee comparisons feasible for large-scale testing programs. Man, Chen, and Li present a novel framework integrating facial expression data with response patterns to enhance cheating detection through distinctive biometric indicators. Together, these complementary methodologies demonstrate how psychometric evidence, statistical classification, computational efficiency, and biometric monitoring can be leveraged to maintain assessment validity while addressing evolving security challenges in digital testing environments. This session provides testing professionals with practical insights for implementing multi-modal security frameworks that balance statistical sophistication with operational feasibility.

Chair:
Huijuan Meng (Amazon Web Services)

Discussant:
James Wollack (University of Wisconsin - Madison)

Presentations:

1. **Jagged Score Profiles and Misconduct Judgments in Language Testing**
   William Belzak, Duolingo

2. **Selecting Decision Thresholds in Random Forest Detection of Test Collusion**
   Irina Grabovsky; Richard Feinberg, NBME

3. **Efficient Answer Similarity Analysis using the M4 Statistic**
   Carol Eckerly, ABIM; Kylie Gorney, Michigan State University

4. **Assessing pre-knowledge cheating via innovative measures**
   Kaiwen Man, University of Alabama; Qipeng Chen, University of Alabama; Jujia Li, University of Alabama

Reimagining Large-Scale Assessment—From Measurement to Meaningful Use (Large Scale Assessment SIGIMIE Session)

Coordinated Paper Session
1:45 PM – 3:00 PM
Intercontinental Los Angeles Downtown, Floor 5 : K-Town

This session brings together four papers that collectively push the boundaries of large-scale assessment (LSA) design, interpretation, and use. The first paper reframes item alignment as a central, micro-level source of validity evidence, arguing for transparent, domain-model-based judgments to improve item quality and fairness. The second paper critiques current approaches to subtest score reporting, proposing a reverse classification tree method to enhance the interpretability and diagnostic value of performance indicators. The third paper introduces a human-centered AI framework that transforms complex process data from assessments like NAEP into actionable student profiles and instructional insights, bridging the long-standing gap between large-scale data and classroom practice. The fourth paper provides an update on the Design-In-Real-Time (DIRTy) approach to personalized assessment, showcasing how AI can support real-time item calibration, alignment, and individualized feedback in adult learning contexts. Together, these papers represent a shift toward more equitable, interpretable, and learner-responsive large-scale assessments. By focusing on alignment, meaningful score use, and the integration of AI in test development, the session highlights emerging models that strengthen the link between assessment design and the real-world needs of educators, learners, and policymakers.

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

Discussant:
Susan Brookhart (Duquesne University)

Presentations:

1. **Advancing Criterion-Based Test Validity Through Improved Item Alignment**
   Marjorie Wine, Accessible Teaching, Learning and Assessment Systems (ATLAS), University of Kansas

2. **Optimizing subscore indicators via reverse classification tree methodology**
   Ji Zeng

3. **Translating LSA Data into Actionable Classroom Insights with AI**
   Hongwen Guo, ETS Research Institute; Matthew Johnson, ETS Research Institute; Jeremy Lee, ETS; Luis Saldivia, ETS; Michelle Worthington, ETS

4. **Design-in-Real-Time Assessments for Adults: Progress on ASAP**
   Stephen Sireci, University of Massachusetts Amherst

The Role of Psychometrics in Higher Education Measurement and Assessment

Organized Discussion
1:45 PM – 3:00 PM
Intercontinental Los Angeles Downtown, Floor 5 : Silver Lake A

Lissette Tolentino, University of Florida; Doris Zahner, Council for Aid to Education; Xiaomei Song, Case Western Reserve University School of Medicine; Elizabeth Smith, Eastern Kentucky University

This organized discussion examines the evolving role of psychometrics and assessment in higher education and the expanding career pathways for psychometricians across university, healthcare, and nonprofit settings. The panelists include Dr. Lissette Tolentino (University of Florida), Dr. Xiaomei Song (Case Western Reserve University), Dr. Doris Zahner (Council for Aid to Education), and Dr. Elizabeth Smith (Eastern Kentucky University). They will address three themes: psychometrics role in higher education measurement and assessment, how politics and artificial intelligence shape assessment practice, governance, and public trust, and opportunities for psychometricians within higher education. The discussion will also center on reframing familiar questions such as, “what are we assessing in higher education and to what end?” by connecting how psychometric theory impacts high-stakes decisions while safeguarding equity and public trust. An interactive dialogue will engage attendees in identifying priority research questions and opportunities for cross-organizational collaboration to advance methodological rigor and practical relevance. The session aims to elevate the importance of psychometrics to higher education measurement and assessment, while offering timely guidance on the competencies necessary to thrive in a rapidly changing landscape.

Chair:
Lissette Tolentino (University of Central Florida)

Classroom Assessment: Validity, Implementation, and Score Reporting/Feedback

Individual Paper Session
1:45 PM – 3:15 PM
Intercontinental Los Angeles Downtown, Floor 7: Roosevelt B

Chair:
Peter Steiner (St.Gallen University of Teacher Education)

Discussant:
Alisha Wackerle-Hollman (University of Minnesota)

Presentations:

1. **Help-Seeking Patterns of Middle School Students in AI-Supported Science Tasks**
   Yi Song, ETS Research Institute; Field Watts, ETS; Lei Liu, ETS; Nichole Jusino del valle, ETS Research Institute; Yun Wang, University of Georgia; Teresa Ober, ETS; Xiaoming Zhai, University of Georgia

   This study examines middle school students’ interactions with an AI agent in a science argument assessment. Using a coding framework of help-seeking requests, we found that students primarily sought clarification and information, with fewer requests involving reflection or feedback, highlighting opportunities to design AI support that fosters deeper scientific reasoning.

2. **Measuring Evidence-based Undergraduate Teaching from Students’ Perspectives: An Interpretation-Use-Consequences Validity Argument**
   Dustin Van Orman, Western Washington University; Greta Moses, Western Washington University; Daniel Hanley, Western Washington University

   Problematic course evaluations in U.S. higher education are widespread. Therefore, we developed a survey to measure students’ experiences of research-based instructional, curricular, and assessment strategies to guide teaching improvements in undergraduate STEM education. This study documents robust evidence across institution types and disciplines to support an interpretation-use-consequences validity argument.

3. **Instructionally Embedded Assessments to Meet Instructional and Summative Uses: Evidence from a Pilot Study**
   Brooke Nash, University of Kansas; Jake Thompson, ATLAS, University of Kansas; Mary Majerus, Missouri Department of Elementary and Secondary Education

   The Pathways for Instructionally Embedded Assessment (PIE) project developed and tested a prototype innovative assessment that delivers instructionally embedded assessments based on cognitive learning models aligned with content standards. In this paper, we describe design features and pilot study evidence to support intended uses according to the theory of action.

4. **Examining the Feedback and Learning Value of Score Reports in U.S. Large-scale Interim Assessment**
   Dustin Van Orman, Western Washington University; Xaviera Gonzalez-Wegener; Chad Gotch, Washington State University; Mary Roduta Roberts, University of Alberta

   Examining the learning value of large-scale assessment score reports is critical to the utility of them, beyond communicating measurement principles. We evaluated representative reports' alignment against research-based feedback principles. Our analysis reveals that the feedback and, therefore, learning value is low in representative reports.

5. **District Scale Up of the Formative Assessment Process**
   John Lane, Michigan Assessment Consortium; Edward Roeber, Michigan Assessment Consortium; Tara Kintz, Michigan Assessment Consortium

   Using period surveys from five districts, this papers sheds light on each district's efforts to scale up the formative assessment process in their schools. In this paper, we report on the frequency and focus of team meetings as well as the programmatic supports district educators used to support their efforts.

NCME Career Achievement Award, Sandip Sinharay: "Reflection on Reporting High-Quality Scores Using Statistics, Measurement, and Professional Judgment"
Invited Session
1:45 PM – 3:15 PM
Intercontinental Los Angeles Downtown, Floor 7: Hollywood Ballroom II

Chair:
Chun Wang (University of Washington)

Discussant:
Ying Cheng (University of Notre Dame) and Scott Monroe (University of Massachusetts, Amherst)

Beverage Break
Social
2:45 PM – 3:45 PM
Intercontinental Los Angeles Downtown, Floor 5 : Wilshire Grand Prefunction

Hot coffee & tea, lemonade, and iced tea will be available.

Inclusive Measurement for Historically Marginalized Groups (IMHM) SIGIMIE Meeting

3:00 PM – 5:15 PM
Intercontinental Los Angeles Downtown, Floor 6: Royal

This SIGIMIE focuses on historically marginalized learner populations and the measurement practices that shape their educational experiences.

We believe that meaningful progress in educational measurement requires sustained attention not only to who is being assessed, but also to how assessments are designed, validated, and interpreted. Marginalization is often intersectional, with overlapping identities amplifying the ways in which students may be excluded or disadvantaged in standardized testing contexts. Through research and dialogue, we strive to develop assessment practices that serve all learners fairly.

Graduate Student eBoards: IRT and DCM

Graduate Student Electronic Board Session
3:30 PM – 4:30 PM
Intercontinental Los Angeles Downtown, Floor 5 : Hancock Park

Presentations:

1. **The Engagement Dip: Examining Effort and Achievement in Low-Stakes Testing**
   Laila Issayeva, James Madison University; Christine DeMars, James Madison University

   This study investigates student disengagement in low-stakes assessments across academic stages. We examined response time effort, self-reported effort, and ability estimates using a 3PL IRT model and an effort-moderated IRT model. Results highlight an engagement dip mid-college and support adjusting for rapid guessing to produce valid proficiency estimates.

2. **Parameter Recovery for Bayesian Testlet Models in Small Sample Sizes**
   Opalhawaye Nyamulani, Fordham University; Leah Feuerstahler, Fordham University

   Testlet models are often developed with respect to moderate to large samples. This study compared the parameter recovery of two testlet models with small sample sizes. Initial findings indicate better recovery for the Bayesian Random Effects testlet model and lower testlet variances. This presentation will also compare other modeling strategies.

3. **Validating a Behavioral Measure of Test-taking Engagement: A Latent Regression Approach**
   Tai Sun Jeong, University of Wisconsin-Madison; James Wollack, University of Wisconsin - Madison

   To validate behavioral measures of test-taking engagement, this study integrated contextual information (e.g., socioeconomic status, and subject-specific anxiety) using a latent regression model with PISA 2022 data. The proposed models improved the precision of individual estimates and provided significant validity evidence, demonstrating a trade-off between overall predictive performance and psychometric explanation.

4. **Model Misspecification Diagnostics for Latent Space Item Response Models**
   Brian Harrold, University of South Carolina; Brian Habing, University of South Carolina

   The recently proposed latent space item response models (LSIRM) extend item response theory to explore local dependence, but diagnostic tools are less explored. Simulation results show that LSIRM geometry is sensitive to model misspecification. Posterior-based diagnostics using LSIRM will be developed and evaluated against established methods to guide model selection.

5. **Many-Facet Rasch Model and Network Analysis: Impact of Rater Effects and Scale Categories**
Daniel Oyeniran, The University of Alabama; Stefanie Wind, The University of Alabama

Accuracy in rater-mediated assessment depends on minimizing rater errors, which may vary by rating categories. We compare Many-Facet Rasch Model and Network Analysis using real and simulated data with different sample sizes, rater effects, designs, and number of categories. Results show interactions among categories, sample size, and rating design.

6. **Disentangling Item and Rater Effects: A Principal Component Analysis Simulation Study**
Wei Huang, Universtiy of Alabama, College of Education; Stefanie Wind, The University of Alabama

We used a simulation to explore principal components analysis (PCA) of standardized residuals to evaluate unidimensionality in Many-Facet Rasch Model analyses. When multiple facets (e.g., students, items, raters) contribute to ratings, the correlation matrix can be constructed in several ways. We evaluate different residual aggregation approaches in this context.

7. **A Copula-Based Joint Model for Item Responses and Response Times**
Sunbeom Kwon, University of Illinois, Urbana-Champaign; Susu Zhang, University of Illinois at Urbana-Champaign

We propose a copula-based joint model for item responses and response times, using a hierarchical framework to model latent ability and speed. By capturing asymmetric and upper-tail dependencies beyond normality assumptions, our approach improves ability estimation accuracy compared with existing methods.

8. **Exploring Response Styles in PISA 2022 Student Questionnaire**
Seung Min Oh, University of Illinois at Urbana-Champaign; Justin Kern, University of Illinois at Urbana-Champaign

This proposal aims to explore response styles in PISA 2022 student assessment. Student cognitive questionnaires are treated as unidimensional despite negatively skewed distributions (OECD, 2024). IRTree models are used to evaluate response styles for subject-specific attributes to provide validity evidence suggested by the Standards (2014).

9. **Effects of Model Misspecification on Abilities in the Presence of Item Complexity**
Yonggi Kim, Chungbuk National University; Wooyeol Lee, Chungbuk National University

Item complexity produces asymmetric item characteristic curves (ICCs), violating the two-parameter logistic (2PL) model's symmetric ICCs. This simulation study quantifies the resulting bias in person parameter estimates and cut-off scores by comparing the misspecified 2PL model against the correctly specified logistic positive exponent (LPE) model that incorporates item complexity.

10. **Investigating Calibration Error in Multidimensional IRT: Implications for Score Accuracy**
Lucy Gitiria, University of Massachusetts, Amherst; Scott Monroe, University of Massachusetts, Amherst

This study examines the effect of calibration error on EAP scores and standard errors for bifactor and testlet multidimensional IRT models. Using the multiple imputation method proposed by Yang et al. (2012), we analyze conditions where calibration error significantly impacts score precision and provide guidance for applied testing practice.

11. **Investigating Heterogeneity in Test Disengagement through IRTree Models with Score-based Partitioning**
Yejin Kim, Seoul National University; Youngin Lee, Seoul National University; Ukil Kim, Seoul National University; Hyun-Jeong Park, Seoul National University

This study aims to detect heterogeneity in test disengagement by analyzing the PISA 2022 mathematics data through the IRTree model with score-based partitioning proposed by Debelack et al. (2025). Results show that there is a significant heterogeneity in students' test disengagement, as well as person covariates associated with it.

12. **Toward a Complete Picture of Effort: Integrating Self-Report with the Disengagement IRTree Model.**
Juste Mehou, James Madison University; Brian Leventhal, James Madison University

Low-stakes assessments are threatened by disengagement, undermining validity. Self-reported effort (SRE) and time-based item-level effort both stem from expectancy-value theory yet reflect distinct motivational mechanisms. This proposal integrates, at the latent level, a measure of self-report with item level time measures of effort to provide a comprehensive account of effort.

13. **Ensuring Classification Consistency Across Grades: Anchor Items in Cognitive Diagnostic Models**
Stephen Tavares, University of Virginia

This study examines how anchor item proportion and placement affect classification consistency in cognitive diagnostic models across grades. Using Monte Carlo simulations and empirical data, we evaluate conditions under which anchor items ensure accurate attribute-level classification, offering guidance for vertical linking in diagnostic assessments.

14. **Data-Driven Estimation of Precedence Relations Among Linear Function Competencies for Formative Assessment**
Peter Steiner, St.Gallen University of Teacher Education; Jan Hochweber, St. Gallen University of Teacher Education, Switzerland; Stephanie Leininger, St. Gallen University of Teacher Education, Switzerland

We explore a data-driven method to establish precedence relations among competencies as a basis for formative assessment using competence structures (e.g., CbKST). Analyzing 547 responses to 41 linear functions items assigned to 84 competencies, we categorized competency pairs by strength of relation and relation direction to identify likely precedence structures.

15. **Developing Adaptive Fit Index Thresholds for Diagnostic Classification Models Using Machine Learning**
Nancy Alila, Department of Educational Psychology, University of Georgia, Athens, GA, USA; Matthew Madison, University of Georgia

Model fit statistics in diagnostic classification models (DCMs) are central to model and Q-matrix selection. The sensitivity of absolute fit indices to sample size necessitates dynamic thresholds over fixed cutoffs. This study leverages machine learning to establish robust, sample-size-invariant criteria, enhancing the validity of DCMs application in real-world settings.

16. **IRT Based Effect Sizes in Meta-Analyses of Intervention Effects**
Jingru Zhang; Qi Huang, Purdue University

Item-level data across studies offer advantages for addressing measurement error in meta-analysis using item response theory (IRT). We show how open-source item-level data and IRT enhance meta-analysis by comparing effect sizes from IRT-based latent scores versus raw sum scores, linking differences to study-level measurement properties.

17. **Double trouble: Assessing RMSEA When Jointly Handling Non-normality and Missing Data**
Yunhang Yin, University of South Carolina; Dexin Shi, University of South Carolina

We investigated six methods for estimating the Root Mean Square Error of Approximation (RMSEA) when both non-normal and missing data occur in the model. Results generally indicated that FIMLZS, MILai, and MIBSL demonstrated robust performance. Key findings, recommendations, and directions for future research are discussed.

Advancing Through-Year Assessments: From Technical Evidence to Interpretation of Summative Scores

Coordinated Paper Session
3:30 PM – 5:00 PM
Intercontinental Los Angeles Downtown, Floor 7: Roosevelt A

For decades, large-scale accountability programs have relied on traditional end-of-year summative assessments to generate state- and district-level data on student learning. While these assessments have provided essential evidence of achievement trends, they often fall short in meeting the immediate instructional needs of teachers and students. The limitations are well known: results arrive months after administration, feedback is typically aggregated to broad reporting categories, and the burden of testing is concentrated into a single high-stakes event. These limitations have fueled interest in innovative models that can both satisfy accountability demands and enhance classroom practice. This session explores through-year assessments as alternative to traditional end-of-year tests. By distributing tasks across the school year, through-year designs provide timely instructional feedback while supporting psychometrically defensible summative scores. Drawing on Montana's statewide MAST assessment, five studies offer a validity framework: policy alignment with stakeholder needs; linking designs that ensure scale reliability; evidence of measurement invariance across time; defensible standard-setting for cumulative scores; and efficient growth modeling without added testing. Together, these studies demonstrate how integrating psychometric rigor with interpretive practices can make through-year assessments both technically defensible and practically valuable.

Discussant:
Susan Lyons (Lyons Assessment Consulting)

Presentations:

1. **From Instruction to Accountability: Policy-Validity Evidence in Montana's MAST Assessment**
Cedar Rose, Montana Office of Public Instruction

2. **Through-Year Assessments: Common Anchor Linking Design**
Duy Pham, New Meridian Corporation; Ourania Rotou, New Meridian Corporation

3. **Through-Year Assessments: Measurement Invariance**
Aaro Douglas, New Meridian Corporation; Andrew Austin, New Meridian Corporation

4. **Through-Year Assessment: Validity Evidence for Standard Setting**
Jami Wireman, New Meridian Corporation; Tim Walker, New Meridian Corporation; Heather Miller, New Meridian Corporation

5. **Through-Year Assessments: Psychometrically Sound and Efficient Approach to Measure Student Growth**
Ourania Rotou, New Meridian Corporation; James Carroll III, New Meridian Corporation

Assessment to Promote Civic Learning: New Volume in the NCME Book Series (Invited Session)

Organized Discussion
3:30 PM – 5:00 PM
Intercontinental Los Angeles Downtown, Floor 5 : K-Town

The role of K-12 schools in cultivating a broad set of civic skills, knowledge, and mindsets has received increasing attention over the past several years, but educators and policymakers lack high-quality measures to inform decisions about how to foster these competencies. A forthcoming edited volume, part of the NCME Educational Measurement and Assessment Book Series, addresses this gap by assembling work from civic scholars and practitioners representing a variety of disciplinary perspectives. In this organized discussion, a volume editor, authors, and a member of the Book Series Editorial Board will discuss major themes from the volume and engage in a conversation with the audience about the need for high-quality, evidence-based approaches to measuring civic outcomes and opportunity to learn.

Chair:
Kadriye Ercikan (ETS Research Institute)

Discussant:
Guillermo Solano-Flores (Stanford University)
Ercikan, Kadriye, ETS Research Institute

Creating and Evaluating Automated Raters

Individual Paper Session
3:30 PM – 5:00 PM
Intercontinental Los Angeles Downtown, Floor 6: Majestic

Chair:
Hacer Karamese (WIDA at the University of Wisconsin-Madison)

Discussant:
Kyoungwon Bishop (University of Wisconsin - Madison)

Presentations:

1. **Measuring Confidence and Consistency in Automated Scoring: Insights from PIRLS**
   Ji Yoon Jung, Boston College; Ummugul Bezirhan, Boston College; Matthias von Davier, Boston College

   We investigated confidence scores and internal inconsistency as quality control metrics for automated scoring (AS). Higher AS confidence aligns with stronger human-machine score agreement, and AS exhibits lower inconsistency than human scoring across and within languages. These metrics help identify challenging responses and guide human-in-the-loop review, improving AS reliability.

2. **Prompt-Engineered LLMs for Engish as A second Language Grammar Diagnosis**
   Doris Abroampah, University of Alberta; Tahereh Firoozi, University of Alberta; Mark Gierl, University of Alberta

   This study integrates prompt-engineered large language models with cognitive diagnostic assessment and Q-matrix coding to identify ESL learners' grammar skill gaps. Using GPT-4o to annotate essays, results show a strong alignment with instructors and revealed diagnostic profiles that support scalable, individualized feedback for language learning and classroom assessment.

3. **Autoscoring Anticlimax: A Meta-analytic Understanding of AI's Short-answer Shortcomings**
   Michael Hardy, Stanford University

   Automated short-answer scoring lags other LLM applications. We meta-analyze 400 results across 40 implementations on ASAP-SAS, modeling Quadratic Weighted Kappa with mixed effects. We show reading items depress performance, decoder-only architectures underperform encoders, and tokenizer size exhibits diminishing returns—patterns traceable to autoregressive training. Findings argue for meaning-grounded, assessment-aligned systems design.

4. **Creating K-12 GenAI Assessment Graders Through Context Engineering**
   Zewei Tian; Lief Esbenshade, University of Washington; Alex Liu, University of Washington; Zachary Zhang, Hensun Innovation; Kevin He, Hensun Innovation; Min Sun, University of Washington

   This research evaluates an LLM grader that uses commercially available foundation models with context and prompt engineering to score student work against a rubric. Compared to human-defined true scores on benchmark datasets, we observed a Quadratic Weighted Kappa of 0.79, suggesting generic foundation models can be effective at scoring.

Data-Intensive Psychometric Research 2: Return to the Item Response Warehouse

Coordinated Paper Session
3:30 PM – 5:00 PM
Intercontinental Los Angeles Downtown, Floor 5 : Ladera Heights

This session highlights research utilizing large volumes of empirical data. Work in educational measurement has conventionally used relatively limited data resources; however, the emergence of a data repository that contains a large volume of standardized and reusable item response data makes it possible to conduct such methodological research at scale in many cases. This resource, the Item Response Warehouse (Domingue et al., 2025), now houses over 1000 item response datasets that are available to researchers via API calls directly from R. These papers—including analysis of prediction quality, the conditional dependencies between discrimination and response time, investigations of interval scales, data augmentation using large language models for psychometric applications, and comparisons of compensatory and noncompensatory models—offer examples of how this data resource can be used to supercharge methods research by examining the performance of techniques across a large volume of empirical data.

Discussant:
Daniel Bolt (University of Wisconsin - Madison)

Presentations:

1. **Precision vs. Prediction: The Relationship between Parametric Standard Errors and Prediction Quality**
   Savira Nadela, Stanford University; Benjamin Domingue, Stanford University

2. **Conditional Dependencies Between Response Time and Item Discrimination: An Item-Level Meta-Analysis**
   Joshua Gilbert; Zach Himmelsbach, Harvard University; Esther Ulitzsch, University of Oslo; Benjamin Domingue, Stanford University

3. **Interval Scales in the Wild**
   Sanford Student, University of Delaware; Wyatt Read, University of Delaware

4. **Large Language Models as Potentially Informative Respondents: Data Augmentation for Psychometric Applications**
   Klint Kanopka, New York University; Austin van Loon, MIT Sloan School of Management; Yuan Huang, New York University; Ruiting Shen, New York University

5. **Compensatory vs. Non-Compensatory MIRT at Scale: Evidence from the Item Response Warehouse**
   Yiqing Liu; Benjamin Domingue, Stanford University

Featured Session: Battle of the Presidents: Past and Future!

Invited Session
3:30 PM – 5:00 PM
Intercontinental Los Angeles Downtown, Floor 5 : Westwood

In this session, 6 Past-Presidents of NCME, and 6 NCME graduate student members square off to respond to (mostly serious) questions on topics in modern educational research and psychometrics. The session promises to be both educational and fun. Come to the session and cheer on the past—or the future!

Chair:
Sara Finney (James Madison University)

Team Members:
Randy Bennett, Assessment Innovation Matters
Richard Patz, University of California, Berkeley
Mark Reckase, Psychometric Solutions
Stephen Sireci, University of Massachusetts Amherst
Ye Tong, National Board of Medical Examiners
Henry Makinde
Valerie Ofori Aboah, The Ohio State University
Alexis Oakley
Khem Sedhai, University at Albany, SUNY
Autumn Wild, James Madison University

Judges:
Susan Davis-Becker, ACS Ventures, LLC
Mark Hansen
Won-Chan Lee, University of Iowa

Scorekeeper:
Haneul Lee, University of Massachusetts Amherst

Featured Session: NCME Task Force Conversation

Invited Session
3:30 PM – 5:00 PM
Intercontinental Los Angeles Downtown, Floor 5 : Wilshire Grand Ballroom III

In the current political climate it is challenging to stay informed about meaningful actions that one can take to protect civil rights. We invite you to join members of the NCME Task Force on Educational Measurement and Civil Rights, who will present a brief overview and then host a scenario-based workshop to bring the work and findings of the task force to life. The session will end with time for participants to draft concrete recommendations for NCME, their own institutions, or policy bodies. These recommendations will be compiled into a “Next Steps” memo to be shared with the full Task Force and the NCME Board.

Chair:
Natalie Rambis (Menlo Education Research)

Panelists:
Britte Cheng, Menlo Education Research
Sarah Beach, University of Virginia
Anne Davidson, CrescendoEd LLC
Magdalen Beiting-Parrish
Carlos Chavez
Howard Everson, City University of New York

Licensure and Certification Research

Individual Paper Session
3:30 PM – 5:00 PM
Intercontinental Los Angeles Downtown, Floor 7: Roosevelt B

Chair:
Nick Trout (Michigan State University)

Discussant:
Daniel Jurich (National Board of Medical Examiners)

Presentations:

1. **Analysis of Repeat Examinee Ability Changes using Multilevel Mixed Model**
   Jing Miao; Shu-chuan Kao, National Council of State Boards of Nursing

   The proposed study uses multilevel mixed modeling to analyze score change patterns among repeat examinees of a licensure exam. It examines time interval between attempts, educational background, and demographics to establish a data-informed threshold for abnormal score gain , supporting test security screening.

2. **Interpretable Features for Response Time Prediction in Medical Assessments**
   Mubarak Mojoyinola; Ye Lin, Ascend Learning; Kari Hodge, Ascend Learning

   Response time is a key metric for evaluating test item quality. Machine learning models can enhance response time prediction by leveraging linguistic features and word embeddings. XGBoost with linguistic and metadata features, achieved 90% accuracy in predicting item response time, demonstrating that interpretable features can match embedding performance.

3. **Advancing Longitudinal Assessment for Medical Continuing Certification: Bayesian Knowledge Tracing Approach**
   Youngjun Lee, The American Board of Anesthesiology; Susan Hibbard, The American Board of Anesthesiology

   This study applies Bayesian Knowledge Tracing (BKT) to longitudinal assessment datasets. We evaluate predictive accuracy, calibration, mastery trajectories, and parameter interpretability across domains and years. Findings show BKT complements existing frameworks by modeling knowledge growth, strengthening validity evidence, and enhancing individualized feedback for medical continuing certification.

4. **Pairing Neural Networks and Word Embeddings to Score Short Answer Medical Items**
   Yooyoung Park, National Board of Osteopathic Medical Examiners; Xia Mao

   The study applies neural network models paired with word embeddings to score short-answer responses on medical exams. Accuracy, matching rate improvements, and overall scoring efficiency are evaluated. The algorithm's benefits and drawbacks are considered in comparison to baseline methods, including exact text matching and k-nearest neighbors.

   Detecting Item Parameter Drift in Longitudinal Assessment for Medical Recertification
   Seongeun Kim, National Commission on Certification of Physician Assistants; Christiana Akande, National Commission on Certification of Physician Assistants; Yanlin Jiang, NCCPA; Nikole Gregg, National Commission on Certification of Physician Assistants

   Many medical recertifications are administered as longitudinal assessments, which offer flexibility in administration and support lifelong learning. Despite these benefits, concerns about item exposure rates arise, which can increase the likelihood of item parameter drift. The current study investigates and provides recommendations for detecting item parameter drift in longitudinal assessments.

Practical Applications of Artificial Intelligence in the Development of Large-Scale Assessments

Coordinated Poster Session
3:30 PM – 5:00 PM
Intercontinental Los Angeles Downtown, Floor 7: Hollywood Ballroom II

Artificial intelligence (AI) is rapidly transforming the development, scoring, and interpretation of large-scale assessments. Assessment programs are increasingly adopting AI to improve efficiency, control, and security, while also enhancing accessibility, personalization, and outcome measurement. Despite its potential, integrating AI into large-scale assessment presents challenges, particularly in meeting longstanding standards for validity, reliability, and fairness. The Standards for Educational and Psychological Testing (AERA, APA, NCME, 2014) emphasize these foundational principles, which remain essential as AI-driven methods evolve. While AI introduces new capabilities, its use must be supported by research and evidence to ensure appropriate and technically sound implementation. Five practical applications of AI in assessment development are presented, each demonstrating how innovation can coexist with best practices. The session explores how emerging AI practices align with established standards of validity and fairness. Presentations include: (1) Using Automated Item Generation to Create Items for Large-Scale Assessments; (2) Scoring Open-Ended Assessment Responses Across Content Areas with Automated Essay Scoring; (3) Using AI to Generate PLDs Aligned with Curricular Standards and Test Specifications; (4) Using AI-Driven Data Augmentation to Improve Training Materials for AES Procedures; and (5) Evaluating AI-Assisted Test Development Procedures with Respect to Validity and Fairness.

Chair:
Anthony Fina

Presentations:

1. **Using Automated Item Generation to Create Items for Large-Scale Assessments**
   Yen Vo, The University of Iowa; Dan Song, The University of Iowa

2. **Scoring Open-Ended Assessment Responses Across Content Areas with Automated Essay Scoring**
   Jing Ma, University of Iowa; Anthony Fina, University of Iowa; Junhee Park, University of Iowa

3. **Using AI-Driven Data Augmentation to Improve Training Materials for AES Procedures**
   Bui Nhat Anh Vu, University of Iowa; Catherine Welch, University of Iowa; Edward Wolfe, Iowa Testing Programs / University of Iowa

4. **Evaluating AI-Assisted Test Development Procedures with Respect to Validity and Fairness**
   Stephen Dunbar, University of Iowa; Catherine Welch, University of Iowa

Practical Issues in Psychometrics

Individual Paper Session
3:30 PM – 5:00 PM
Intercontinental Los Angeles Downtown, Floor 5 : Silver Lake B

Chair:
Liru Zhang (Consultant)

Discussant:
Aaron Myers (American Board of Internal Medicine)

Presentations:

1. **Testing Item-level Heterogenous Treatment Effect with Penalized DIF Detection—Do Effects Disappear?**
   Rujun Xu, University of Virginia; James Soland, University of Virginia

   When measuring treatment effects in intervention studies, previous researchers found that some items show larger effects than others. While it could be that the intervention affects items in different ways, we try to investigate other possibilities in this paper, such as alpha inflation, improper randomization, and intersectional forms of bias.

2. **Detecting Rater Effects in Human Scores: A Cumulative Approach Using Statistical Indicators**
   Yangmeng Xu, Pearson; Edward Wolfe, Iowa Testing Programs / University of Iowa

   This study explored a cumulative approach for detecting rater effects (RE) using statistical indicators and compared the effectiveness of these indicators by evaluating how quickly (i.e., the number of student responses needed) and how accurately (i.e., Type I and II error rates) they detect RE among human scores.

3. **Building a Better Rater: Using Confusion Matrixes and Assembly Rules to Improve Performance Assessors**
   Andrew Krumm; Russell Almond, Florida State University; Kayla Marcotte, University of Michigan; Blaire Carpenter, Florida State University; Jiawei Li, Florida State University

   This paper applies confusion matrixes to understand rater accuracy as well as rater-focused assembly rules for populating individual matrixes. Using a large corpus of performance assessments from surgery education, this paper demonstrates how to calculate various rater severity metrics.

4. **Modeling Accuracy in Oral Reading Fluency Assessment: Exploring Count Data Modeling Options**
   Paul Foster, Southern Methodist University; Yusuf Kara, University of Miami; Joseph Nese, University of Oregon; Akihito Kamata

   Oral reading fluency incorporates fast and accurate reading and is an important indicator of overall reading competency. Accuracy is measured through the total word-level reading success or failure. This study explores the feasibility of various count data models for more realistic measurement of accurate reading.

Practical Issues with Process Data

Individual Paper Session
3:30 PM – 5:00 PM
Intercontinental Los Angeles Downtown, Floor 5 : Silver Lake A

Chair:
Fusun Sahin (Curriculum Associates)

Discussant:
Hong Jiao (University of Maryland)

Presentations:

1. **To skip or not to skip: Does accidental item skipping predict performance?**
   Ella Anghel, Ben-Gurion University in the Negev; Matthias von Davier, Boston College

   ePIRLS includes messages to prevent test-takers from accidentally leaving items unanswered, but they could demotivate or distract participants. To test this, we compared the performance of 111 test-takers who skipped an item accidentally to matching non-skippers. We found that viewing the message resulted in lower item scores (OR = 0.56).

2. **Digital Reading, Analog Foundations: Reading Literacy in the Digital Age**
   Jeneve Swaby, Boston College; Matthias von Davier, Boston College

   Do digital reading assessments measure something meaningfully distinct from print? Using PIRLS/ePIRLS 2016, we implement survey-weighted regressions pooled across plausible values to evaluate construct distinctiveness conditional on print proficiency, covariates, and process indicators. Findings show digital reading taps the same core literacy with small engagement effects.

3. **Standardizing Predictors for Effect Sizes in Multilevel Models**
   Guanyu Chen, The University of British Columbia; Amery Wu, University of British Columbia; Yan Liu, Carleton University

   Effect sizes are essential in multilevel modeling, yet computation becomes challenging when random slopes induce heteroscedasticity. We propose a new standardization approach that centers and scales predictors, simplifying variance decomposition and yielding equivalent effect size measures as Johnson's mixture approach. The method is practical, interpretable, and accessible for applied researchers.

4. **Enhancing Prediction Accuracy of Practice Tests for Multi-Subject Certification Exams**
   Yung-Chen Hsu, GED Testing Service; Tsung-hsun Tsai, Research League, LLC

   This study evaluates predictive models linking practice test scores to certification outcomes. The challenge arises from the differing expectations of subsidizers, trainers, and testing organizations, further complicated by diverse test taker motivations. We employed machine learning techniques alongside outlier removal to enhance model accuracy and identified a balanced solution.

5. **Effect Size Estimation for Multilevel Models in Complex Survey Designs**
   Guanyu Chen, The University of British Columbia; Amery Wu, University of British Columbia; Yan Liu, Carleton University

   This study introduces two methods, weighted mixture and weighted standardization, for estimating effect sizes in multilevel models that include random slopes and sampling weights, which often make effect size calculations challenging. With 2018 PISA data, the research highlights the importance of sampling weights in complex survey data and accessible solutions.

Raters and Rating Scales

Individual Paper Session
3:30 PM – 5:00 PM
Intercontinental Los Angeles Downtown, Floor 5 : Boyle Heights

Chair:
HyunJoo Jung (Inha University)

Discussant:
HyunJoo Jung (Inha University)

Presentations:

1. **Rating Time Predicts Illusory Halo: A Mixture Rasch Facets Analysis**
   Kuan-Yu Jin, Hong Kong Examinations and Assessment Authority; Thomas Eckes, University of Bochum

   Halo effects refer to a persistent rater error posing a significant threat to the validity and fairness of assessments involving human raters. This research introduces a novel approach to analyzing halo effects by considering rating times as an additional data source. A Chinese writing assessment was analyzed for demonstration.

2. **Practical Impacts of Within-Subgroup Rating Scale Malfunctioning for Polytomous IRT Models**
   Stefanie Wind, The University of Alabama; Theode Niyirinda, University of Alabama

   We used a real data analysis and simulation to examine the effects of rating scale mal-functioning (RSMF; i.e., disordered, non-distinct, or non-invariant thresholds) within a test-taker subgroup on psychometric quality using different polytomous IRT models. We observed only small effects for within-subgroup RSMF on overall indicators of psychometric quality.

3. **Exploring the Impact of Differences in Category Use between Human and Automated Raters**
   Stefanie Wind, The University of Alabama; Wei Huang, Universtiy of Alabama, College of Education

   We considered the impact of differences in scoring patterns between human and AI raters on the accuracy of rater effect indicators. Our results suggest that differences in category use, which may be masked in correlations between rater types, can impact the accuracy of rater effect detection for both rater types.

4. **Local Independence and its Impact on Rater Effect Indices in the MFRM**
   Mirai Nagasawa; Stefanie Wind, The University of Alabama

   We explored the practical consequences of dependence between raters on rating quality indicators based on the MFRM. Results indicated that the influence of rater dependence on the targeted indices was minimal, suggesting that the MFRM is robust to the presence of rater dependence.

5. **Evaluating Validity-Based Anchor Responses as an External Criterion for Rater Accuracy**
   Yue Huang, Measurement Incorporated; Corey Palermo, Measurement Incorporated; Yong He, Measurement Incorporated; Troy Chen, Measurement Incorporated

   We simulated rater–response scoring using item response theory (IRT) models to compare validity-based and conventional inter-rater reliability (IRR). Results show rater–true agreement consistently exceeds inter-rater agreement, with the largest divergence with low-discrimination, high-difficulty responses. Validity responses provide a more accurate benchmark, highlighting limitations of consensus-based IRR.

Test Security SIGIMIE

Meeting
5:30 PM – 6:30 PM
Intercontinental Los Angeles Downtown, Floor 6: Royal

This SIGIMIE serves as both a forum and a platform for practitioners and scholars to share insights, exchange ideas, and support each other.

This collective effort aims to mitigate the risks faced by testing companies and organizations. Our goal is to enhance NCME's involvement in this vital research area by disseminating findings, promoting innovative techniques, and developing practical approaches applicable across various assessment products.

NCME Fitness Run/Walk
Invited Session
6:00 AM – 7:30 AM
Offsite: Lobby of NCME Hotel (InterContinental LA Downtown)

The fun run is a 2.5K or 5K walk or run. All abilities are welcome. Participants will meet at the lobby of the NCME Hotel (InterContinental Los Angeles Downtown --900 Wilshire Blvd, Los Angeles, CA 90017) at 6 AM. We will take a chartered bus to Elysian Park for the event and return via bus as well. Registration includes a shirt that can be picked up at the NCME Hotel Information Desk. Contact Katherine Castellano (KEcastellano@ets.org) with any questions.

Mother's Room
Meeting
7:30 AM – 5:00 PM
Intercontinental Los Angeles Downtown, Floor 5 : Glassell Park

Private mother's room for nursing.

NCMEntoring Program
Meeting
7:30 AM – 5:00 PM
Intercontinental Los Angeles Downtown, Floor 6: Roxy

Launched at the 2016 annual meeting, the NCMEntoring Program aims to support the transition of graduate student members and recent graduate members from their graduate programs to professional careers. Early professionals (mentees) are paired with members (mentors) experienced in NCME-related fields: psychometrics, assessment, certification, evaluation, and other aspects of educational measurement.

This experience offers mentees the opportunity to explore possible career paths and/or research interests and for mentors to support the development of potential colleagues and contribute to the field. Each year, over 100 NCME members participate in the NCMEntoring Program and participant feedback has been positive. The Program hopes to cultivate long-term relationships between mentors and mentees.

Low Sensory Room
Meeting
7:30 AM – 6:00 PM
Intercontinental Los Angeles Downtown, Floor 6: Jade

Attendees who need areas that are quiet with reduced light and noise may take comfort in the low-sensory room.

ASA Listening Session with NCME Members: Re-Envisioning the National Center for Education Statistics

Organized Discussion
7:45 AM – 9:15 AM
Intercontinental Los Angeles Downtown, Floor 5 : K-Town

Session Description: The American Statistical Association (ASA) is engaged in a project to reimagine and modernize the National Center for Education Statistics (NCES) to be more effective, efficient, and responsive to the needs of the research and education communities. This listening session provides NCME members with an opportunity to share their experiences and insights directly with ASA. Facilitated by Michelle Crosby (ASA), the conversation will be structured around three themes: (1) how NCME members intersect with the work of NCES—as data users, contractors, technical advisors, or in other capacities; (2) how recent federal changes have affected members' work and the broader measurement community; and (3) members' ideas for how a reimagined NCES could better serve the field. Input gathered during this session will inform ASA's recommendations for the future of NCES. All NCME members are welcome to attend and contribute.

Chair:
Susan Lyons (Lyons Assessment Consulting)

Presenter(s):
Michelle Crosby, American Statistical Association

Discussant:
Carr, Peggy - Former Commissioner, National Center for Education Statistics, U.S. Department of Education

Alternate and Personalized Assessment

Individual Paper Session
7:45 AM – 9:15 AM
Intercontinental Los Angeles Downtown, Floor 7: Roosevelt A

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

Discussant:
Yi-Chen Wu (National Center on Educational Outcomes, University of Minnesota)

Presentations:

1. **Examining Variation in Alternate English Language Proficiency and Alternate Content Assessment Outcomes**
   Glenn Poole, WIDA; Narék Sahakyan, University of Wisconsin - Madison / WIDA

   We examine the academic outcomes of English Learners with the most significant cognitive disabilities. We discuss extensions to methods for examining relationships between English language proficiency and alternate content assessment outcomes, used to support setting reclassification criteria for these students, and present evidence on important variation within and between states.

2. **Assessing Young Learners: Building Fairer and More Accurate Reporting Structures**
   Kerry Englert, Seneca Consulting, LLC; Pohai Kukea Shultz, University of Hawaii - Manoa

   As education systems across the nation utilize kindergarten readiness assessments, we must ensure they accurately reflect the abilities of all students. This presentation highlights an assessment project that aims to provide educators and parents with reliable, fair, and actionable data on their students' skills, knowledge, and language readiness.

3. **A Quasi-Experimental Approach to Studying Reclassification Policy for English Learners**
   Hirotaka Fukuhara, Pearson; Dipendra Subedi, Pearson Assessments; Scott Strickman, Pearson

   This study empirically validates the reclassification cut score for an English proficiency assessment for English learners using regression discontinuity design. Findings confirm the cut score's appropriateness. By linking proficiency assessment to subsequent achievement, the research advances outcome-based validation practices, offering a replicable model for policy-relevant measurement in multilingual educational contexts.

4. **Psychometric Implications of Innovative Candidate-Centered Educational Assessments**
   Jacquelyn Thompson Torbet; Alison Linder, Pearson

   This research identifies multiple candidate-focused assessment practices being implemented in professional certification assessments that address the emerging need for personalized assessment and that challenge traditional standardization. The research underscores the importance of balancing innovation with rigor to ensure future assessments benefit both individuals and society.

Detecting and Assessing DIF

Individual Paper Session
7:45 AM – 9:15 AM
Intercontinental Los Angeles Downtown, Floor 5 : Silver Lake B

Chair:
Ru Lu (Educational Testing Service)

Discussant:
Moses Mohamed (University of South Florida)

Presentations:

1. **Comparing DIF detection methods for operational CAT items**
   Zachary Mayne, IXL; Michael Peabody, IXL Learning; Yu Zhao

   Most differential item functioning (DIF) detection methods were designed for use with fixed-form assessments. Those who designed methods for computerized-adaptive testing (CAT) did so for use with pretest items and sparse data. This study compares CAT DIF detection methods using operational CAT items.

2. **DIF Detection Under the Compensatory Reparameterized Unified Model (C-RUM)**
   Kevin Krost

   The power and Type I error rates of Lord's Wald test and the Likelihood Ratio (LR) test when using the compensatory reparameterized unified model (C-RUM). DIF magnitude had the largest effect, and DIF detection method had a negligible effect. Sample size, DIF type, and Q-matrix complexity had large effects.

3. **Evaluating Average Signed Area as a DIF Index in DIF-Free-Then-DIF Strategy**
   Wei-Chia Su, National Sun Yat-sen University; Ching-Lin Shih, National Sun Yat-sen University

   The logistic regression, combined with the DIF-free-then-DIF strategy, was used to simultaneously assess uniform and nonuniform DIF in this study. Through a series of simulation studies, the Average Signed Area (ASA) seem suitable only for uniform DIF assessment, yet seems limited for nonuniform DIF assessment.

4. **Three residual-based DIF assessment strategies for assessing uniform and nonuniform DIF items.**
   Hu Po Hsien, National Sun Yat-sen University; Ching-Lin Shih, National Sun Yat-sen University

   This study compared three residual-based DIF methods, scale purification with removed DIF items one-at-a-time, scale purification with removed DIF items all-at-a-time and the two-stage strategy with an eight-item matching variable. The performance of these method were evaluated through a series of simulation studies and the findings and implications were discussed.

5. **Detecting Global and Net DIF in Polytomous Items Using RDIF**
   HyunJoo Jung, Inha University; Hwanggyu Lim, Inha University; Jaime Malatesta, GMAC; Yongsang Lee, Inha University

   This study aims to extend the recently introduced residual DIF detection framework (Lim et al., 2022), originally developed for dichotomous items, to polytomous IRT models. To evaluate the effectiveness of the proposed methods in identifying items with DIF, we conduct both a simulation study and an empirical analysis.

Innovations in Assessment Development

Individual Paper Session
7:45 AM – 9:15 AM
Intercontinental Los Angeles Downtown, Floor 5 : Westwood

Chair:
Nathan Dadey (The National Center for the Improvement of Educational Assessment)

Discussant:
Nathan Dadey (The National Center for the Improvement of Educational Assessment)

Presentations:

1. **Optimizing Block Assembly in Forced-Choice Assessments**
   Sirui Wu, University of British Columbia; Amery Wu, University of British Columbia

   This study evaluates new optimization functions for assembling forced-choice test blocks that balance both option discrimination and preference. Simulations across trait distributions and correlations demonstrate that including preference, long overlooked in modern practice, improves ranking accuracy in fixed-form tests compared to conventional discrimination-only assembly methods.

2. **Detection of Enemy Items on K-12 Science Assessment Using Natural Language Processing**
   Mina Lee, Cambium Assessment; Seonho Shin, Cambium Assessment

   This study investigates term-frequency-based cosine similarity to detect enemy items in K–12 science assessments. Analysis of a large item bank indicates that higher similarity predicts enemy relationships, particularly for items sharing scientific phenomena or keywords. The findings emphasize the importance of establishing similarity thresholds to guide subject matter expert review.

3. **Addressing Dyslexia Screening Floor Effects: An Adaptive Timing Approach Inspired by Psychophysics**
   Wanjing (Anya) Ma, Stanford University; Benjamin Domingue, Stanford University; Jason Yeatman, Stanford University

   This study addresses floor effects in measuring single-word recognition, a key indicator of dyslexia risk. We propose an adaptive timing approach: students start with unlimited-time items and progress to fast-timed items based on reading ability. This enhances the testing experience for beginning readers while preserving score sensitivity for all abilities.

4. **Construction of Instructionally Sensitive Items: Predicting Math Items' Sensitivity Using Item Properties**
   Alexander Naumann; Stephan Schönenberger, St. Gallen University of Teacher Education, Switzerland; Stephanie Leininger, St. Gallen University of Teacher Education, Switzerland; Marit List, DIPF | Leibniz Institute for Research and Information in Education, Germany; Jan Hochweber, St. Gallen University of Teacher Education, Switzerland; Johannes Hartig, DIPF

   Valid inferences on teaching drawn from students' test scores require that tests and items are instructionally sensitive. However, only little is known about the purposeful construction of instructionally sensitive test items. Thus, we conducted an item experiment in math aiming at identifying item properties that contribute to items' instructional sensitivity.

Randomly Parallel Testing: New Methods, Emerging Questions, and Practical Solutions

Coordinated Paper Session
7:45 AM – 9:15 AM
Intercontinental Los Angeles Downtown, Floor 5 : Ladera Heights

This symposium explores Lord's 1955 innovation, Randomly Parallel Testing (RPT). Research shows RPT reduces cheating, improves score generalizability, and generates unique, equivalent test forms. Five presentations examine RPT from different perspectives, highlighting the emerging role of RPT as both a theoretical innovation and a practical solution to today's testing challenges. The opening paper frames RPT as a transformative departure from traditional fixed-form testing, emphasizing its potential to reduce cheating, improve defensibility, and better align with modern technological realities. Complementing this broad vision, another study examines IRT-based equating strategies for score comparability across RPT and conventional forms, offering simulation and real-data evidence. Another presentation explores the use of AI, particularly ChatGPT, in generating SmartItems that provide the item pools required for RPT, focusing on quality, difficulty calibration, and fraud reduction. A separate contribution addresses the limitations of applying factor analysis within RPT contexts, proposing new statistical checks and simulation-based strategies for recovering item pool structures. Finally, the symposium concludes with research on classification indices under RPT, advancing methods to evaluate mastery decisions and reliability in criterion-referenced settings. Collectively, these papers illustrate the promise and complexity of implementing RPT while charting practical paths for its adoption in operational assessments.

Discussant:
Alina von Davier (Duolingo)

Presentations:

1. **From Tradition to Transformation: The Case for Randomly Parallel Testing**
   Robert Brennan, The University of Iowa; David Foster

2. **IRT Equating for Randomly Parallel Testing**
   Seungwon Shin, University of Iowa; Qiao Liu, UNC Charlotte; Won-Chan Lee, University of Iowa; Stella Kim, University of North Carolina Charlotte

3. **Examining the Effectiveness of AI-assisted Creation of SmartItems**
   Peter Tran, University of Massachusetts, Amherst; Craig Wells, University of Massachusetts Amherst; David Foster; Stephen Sireci, University of Massachusetts Amherst

4. **Factor Analysis in Randomly Parallel Tests: Limitations and Emerging Solutions**
   Sergio Araneda, Caveon Test Security

5. **Proficiency Scoring with AI-Predicted Item Parameters via Iterative Parameter Updating**
   Won-Chan Lee, University of Iowa; Stella Kim, University of North Carolina Charlotte; Yeonju Lee, University of Iowa; Seungwon Shin, University of Iowa

Score Comparability Challenges in Applied Assessment Contexts (Contemporary Issues in Scaling, Linking & Equating SIGIMIE Session)

Coordinated Paper Session
7:45 AM – 9:15 AM
Intercontinental Los Angeles Downtown, Floor 5 : Silver Lake A

Ongoing challenges in assessment include efforts to redesign assessments while maintaining score comparability. This session presents several examples of these challenges and approaches to addressing them. The first presentation examines concordance tables for college admissions tests, highlighting concerns about their accuracy when constructs, testing populations, and subpopulations differ or change across the concorded tests. The second presentation focuses on linking a licensure test before and after a transition to reduce and integrate assessment components, including analyses of linkings, calibrations, and performance evaluation, item difficulty and comparability. The third presentation covers a K-8 exam's transition from a two-step testing (one fixed-form and one multi-stage) to an integrated multi-stage test, including efforts to ensure comparability in assessment performance, reliability, and (differential) item functioning. The fourth presentation explores comparability challenges arising from automated procedures for generating test items and their scoring parameters, and the use of automated, non-human scoring for open ended responses. The fifth presentation describes empirical evaluations of digital versus paper-based administrations of NAEP, including mode effects across grades and subjects, exam results that account for mode effects, and self-reported perceptions. These papers provide new insights for supporting score comparability in the recently and rapidly changing assessment field.

Chair:
Tim Moses

Presentations:

1. **College Admissions Issues Related to Linking and Equating**
   Kurt Geisinger

2. **1+1>2: Linking Separate Knowledge and Skills Examinations to An Integrated Examination**
   Hao Song, Association of State and Provincial Psychology Boards

3. **Evaluating Score Comparability of a Redesigned Multistage Assessment**
   Shirley Li, Savvas Learning Company; Jinnie Choi, Savvas Learning Company; Rongchun Zhu, Savvas Learning Company

4. **AI Applications for Assessment Automation: Implications for Score Comparability**
   Tim Moses, Buros Center for Testing; Jinghua Liu, The National Board of Osteopathic Medical Examiners; Seongeun Kim, National Commission on Certification of Physician Assistants; Nick Trout, Michigan State University

5. **Digital Assessment Mode Effects in K-12: Insights from Large-Scale NAEP Data in Reading, Mathematics, Science, and Social Science**
   Paul Jewsbury, ETS; Yue Jia, ETS Institute; Nuo Xi, ETS; Meng Wu, Educational Testing Service; Adrienne Geijer, ETS Research Institute; xueli xu, ETS Institute; Carol Eckerly, ABIM; John Donoghue, ets

Scoring Early Literacy Tasks: Cross-Vendor Research and Perspectives

Coordinated Paper Session
7:45 AM – 9:15 AM
Intercontinental Los Angeles Downtown, Floor 6: Majestic

Early literacy assessments are becoming a critical piece of K-12 large scale assessment to support evidence-based reading instruction (Brunetti et al., 2025). Because critical decisions are made using scores from these assessments, it is essential that they are rigorously evaluated and that we are transparent about the methods used in the design, human scoring, automated scoring, test scoring, and implementations. This session focuses on key elements in this process: human scoring methods for a broad range of early literacy items, AI modeling of those items, and the impact of scoring approaches on classification decisions. It also focuses on feasibility and instructional value using teacher interviews, examination of process data, and observations of students. The work is demonstrated across three early literacy assessment developers (Cambium Assessment, Curriculum Associates, and Edmentum) in order to illustrate the range of methods and outcomes.

Chair:
Susan Lottridge

Discussant:
Joseph Nese (University of Oregon)

Presentations:

1. **Implementing and Evaluating Human Scoring for Early Literacy Tasks**
   Debbie Dugdale, Cambium Assessment, Inc.

2. **Implementing and Evaluating AI Modeling for Early Literacy Tasks**
   Christopher Ormerod, Cambium Assessment

3. **Designing for Authenticity and Accuracy: UX Research for Voice Recognition Early Literacy Tasks**
   Danielle Schwartz, Curriculum Associates

4. **Evaluating Authenticity and Accuracy for Voice Recognition Early Literacy Tasks**
   Aimee Boyd, Curriculum Associates

5. **Evaluating Oral Fluency Tasks for Early Literacy Composite Score Development**
   Catherine Oberle, Edmentum, Inc.; Yi He, Edmentum

The Revision of the Testing Standards: Operations (Invited Joint AERA/NCME Session)

Organized Discussion
7:45 AM – 9:15 AM
Intercontinental Los Angeles Downtown, Floor 5 : Wilshire Grand Ballroom III

This is the second of three joint sessions designed to update the memberships of AERA and NCME and invite input on the revision of the Standards for Educational and Psychological Testing. This session will focus on the Operations chapters, with particular attention to test development, test administration, scoring, rights and responsibilities of test takers and test users, and documentations for tests.
The Joint Committee (JC) has been working on revising the 2014 Standards. In this second session, participants will receive an update on the revision process and learn about major changes proposed for the opeations chapters. Most importantly, the session will provide time to hear directly from participants—their reactions to the updates as well as their questions, hopes, and concerns about the ongoing revision.

Chair:
Ye Tong (National Board of Medical Examiners)

Discussant:
Andy de los Reyes

Presenter(s):
Ye Tong, National Board of Medical Examiners; Mark Wilson, University of California, Berkeley; Qiwei He, Georgetown University; Rochelle Michel; Cara Laitusis, Center for Assessment; Maria Marquine; Stephen Stark, University of South Florida; Andy de los Reyes; Michael Rodriguez, University of Minnesota

Visualizing and Interpreting DIF

Individual Paper Session
7:45 AM – 9:15 AM
Intercontinental Los Angeles Downtown, Floor 7: Roosevelt B

Chair:
Kelli Treadwell

Discussant:
Irina Grabovsky

Presentations:

1. **Rethinking Item Fairness Using Single World Intervention Graphs**
   Youmi Suk, Teachers College, Columbia University; Weicong Lyu, University of Macau

   This study proposes a causal framework for assessing item fairness using single world intervention graphs. We define causal DIF as the difference in item functioning between two hypothetical worlds, and demonstrate its advantages over traditional DIF methods through a simulation study and a real data application.

2. **Beyond DIF Detection: A Downstream Clustering Framework for Small Sample Parameter Estimation**
   Yale Quan, University of Washington; Chun Wang, University of Washington

   This study introduces the Transitive DIF Clustering (TDC) algorithm, a graph-based algorithm for determining measurement invariance. By leveraging DIF probabilities, TDC reframes invariance as a clustering problem. Through simulations we demonstrate that TDC is better able to recover the true measurement invariance structure as compared to traditional clustering algorithms.

3. **Multidimensional Rasch Model and Rasch Tree Analysis of Cognitive Load Rating Scale**
   Christoper Ocheni, The University of Alabama, Tuscaloosa; Daniel Oyeniran, The University of Alabama; Justice Dadzie, The University of Alabama

   This study evaluated the Cognitive Load Rating Scale using multidimensional Rasch modeling and Rasch tree analysis. Results showed that the multidimensional Rasch model, with acceptable psychometric properties, significantly fits the data better than the unidimensional model. Rasch tree analysis revealed differential item functioning by age and gender, confirming construct validity.

4. **Guidelines for the Interpretation of NCDIF as an Effect Size Measure**
   Trung Le, University of Illinois Urbana-Champaign; Víctor H Cervantes, University of Illinois Urbana-Champaign

   We present a systematic framework for developing meaningful interpretation guidelines for NCDIF. To achieve this, we investigate its comparability with the Delta Mantel-Haenszel (ΔMH), examine one of the NCDIF estimators to evaluate the accuracy of the derived classification rules, and identify an approximate bias correction for this estimator.

5. **Multidimensional dMACS for Practical DIF Effect Sizes**
   Hao Zhou; Yu Cui, Dongguan City University

   MdMACS extends dMACS to multidimensional CFA, yielding item-level DIF effect sizes with scalar/metric decomposition and covariance diagnostics. Simulations show high recall with permutation thresholds and controlled false positives; applications demonstrate minimal partial invariance removes practical misfit. Open-source code and reporting templates support adoption.

Quiet Room

Meeting
8:00 AM – 7:00 PM
Intercontinental Los Angeles Downtown, Floor 6: Novelty

During the Annual Meeting, attendees who desire a quiet place to relax or prepare for a presentation may visit quiet rooms available 8:00 am-7:00pm daily.

Graduate Student eBoards: AI & Machine Learning

Graduate Student Electronic Board Session
8:15 AM – 9:15 AM
Intercontinental Los Angeles Downtown, Floor 5 : Hancock Park

Presentations:

- **Improving Students' Code Comprehension Ability Through LLM-Generated Practice Problems (eBoard 1)**
  William Kwako, Georgia Institute of Technology

  Students struggle to meet learning objectives in introductory computer science courses. In particular, students' ability to read code lags behind their ability to write it. In this project, we develop an educational webapp that generates tailored practice problems for students to solve, aimed at strengthening students' code comprehension skills.

- **AI-Augmented Automatic Item Generation for Culturally Responsive Assessments (eBoard 2)**
  Anna Nasyrova, University of Massachusetts, Amherst

  The primary goal of this paper is to inform the discussion on contemporary mechanisms for AI-augmented automatic item generation for culturally responsive reading assessments. We adopt a three-step workflow for culturally responsive items generation. The results include practical suggestions in the context of formative assessments.

- **From Item to Language Models: A Review of AI-Assisted Item Generation (eBoard 3)**
  Meng Lyu

  This literature review explores AI-assisted Automatic Item Generation (AIG), highlighting key findings: the rise of hybrid models combining templates and large language models; the evaluation paradox slowing validation; the emergence of Meta-AI for quality control; and the transformed expert role in human-AI collaborative workflows enabling scalable, high-quality item production.

- **Evaluating AI-generated NAEP Item Quality by Achievement Levels and Prompt Engineering (eBoard 4)**
  Haneul Lee, University of Massachusetts Amherst; Yeonu Kim, Inha University

  This study examines ChatGPT's potential to generate Grade 8 NAEP mathematics items across achievement levels and prompt engineering strategies. The quality of GPT-generated items is evaluated by experts through surveys and focus group interviews. By analyzing evaluation results, this study contributes to establishing methodological foundations for AI-driven assessment design.

- **Robustness of Transformer-Based AES Systems: A Systematic Evaluation with Adversarial Attacks (eBoard 5)**
  Lingchen Kong, University of Florida; Jinnie Shin, University of Florida

  This study evaluates the robustness of five Transformer AES models on PERSUADE 2.0 using perturbations at punctuation, character, word, and sentence levels. Robustness is quantified by score-change rates (increase/same/decrease), complementing quadratic weighted kappa (QWK). DeBERTa shows the strongest rubric alignment, Longformer resists padding yet is typo-brittle, while BERT/DistilBERT remain inflation-prone.

- **AI-Enhanced Multi-Agent Systems for Psychometric Assessment (eBoard 6)**
  Jujia Li, University of Alabama; Kaiwen Man, University of Alabama; Joni Lakin, University of Alabama

  This paper introduces an AI-enhanced multi-agent system (MAS) for psychometric assessment (PA-MAS) that integrates IRT, CDM, process, and qualitative models. By incorporating large language models for think-aloud coding, the system increases automation, fairness monitoring, and process validity and offers scalable and verifiable adaptive education measurement solutions.

- **Improving Writing Instruction with AI-Driven Analytic Scoring and Rubric-Aligned Feedback (eBoard 7)**
  Farzana Yasmin, University of Alberta, Measurement, Evaluation & Data Science (MEDS); Mark Gierl, University of Alberta

  Providing consistent, high-quality feedback at scale is challenging. We propose a rubric-grounded, prompt-based approach using a large language model to generate analytic scores and targeted feedback on ASAP++ essays. With human oversight and calibrated outputs, the API-served model supports formative assessment by delivering faster, consistent, and actionable feedback for use.

- **Evaluating AI Tools in Coding Motivational-Developmental (MD) Constructs from Open-Ended Student Responses (eBoard 8)**
  Olasunkanmi Kehinde, Norfolk State University; Kelvin Afolabi, University of Virginia; Scott Etheridge, Elizabeth City State UNiversity; Cyril Udensi, Elizabeth City State UNiversity; Adedolapo Adebayo, Norfolk State University

  This study evaluates AI models (ChatGPT, Claude, and Gemini) in replicating human scoring of motivational-developmental responses in undergraduates. Comparing accuracy, precision, recall, and F1-scores, results show Gemini consistently outperforms others, highlighting its potential for reliable AI-assisted assessment of non-cognitive attributes in higher education.

- **Predicting when LLMs-assigned scores will be similar to human-assigned scores (eBoard 9)**
  Jae Jun Jong

  Inspired by Heo et al. (2024), I investigated whether LLMs' understood inputs (i.e., hidden states; Duan et al., 2024) are different when LLM-assigned and human-assigned scores are similar and different. Visual and numerical comparisons were conducted. The results show that there are no apparent differences in hidden states.

- **Machine Learning Classification of IRT Parameters and PLDs Using Transformer-Based Texts (eBoard 10)**
  Junhee Park, University of Iowa; Sandra Sweeney, Cognia; Louis Roussos, Cognia

  This study investigates how the item text features from a transformer-based large language model classify the clusters with item discrimination and item difficulty by using four machine learning models. Moreover, it suggests validity evidence of the alignment of item characteristics-to-Performance Level Descriptors from the same item text features.

- **Clustering by Language, Guided by Theory: Using NLP to Support Scale Development (eBoard 11)**
  Josiah Hunsberger, James Madison University; Joshua Shulkin, University of Maryland

  Sample size considerations can limit iterative model development in a CFA framework with narrowly defined populations. Using an existing psychological measure, this study found that natural language processing (NLP) identified factors were comparable to existing factor structures. These results suggest an NLP approach can aid in small sample scale development.

- **Text-Based Approaches to Item Alignment to Content Standards in Reading & Writing Tests (eBoard 12)**
  Yanbin Fu, University of Maryland, College Park; Hong Jiao, University of Maryland; Tianyi Zhou, University of Maryland; Nan Zhang, University of Maryland; Ming Li, University of Maryland; Qingshu Xu, University of Maryland, College Park; Sydney Peters, University of Maryland, College Park; Robert Lissitz, University of Maryland, College Park

  Aligning test items to content standards supports content-based validity. Using SAT/PSAT reading–writing items, this study fine-tuned small language models for domain and skill alignment and examine input-field and sample-size variations. Models perform strongly. Analyses of the embedding-based similarity clarify systematic misclassifications among closely related skills.

- **Auditing AI Rationales as Validity Evidence for Standard Setting Cut Score Decisions (eBoard 13)**
  Claudia Ventura, University of Connecticut; Joe Grochowalski, College Board; Amy Hendrickson, The College Board

  GenAI was used to address limitations in standard settings, often opaque and hard to audit. We test a GenAI standard-setting by validating its rationales and skill-probabilities against proposed AP and SAT. Results show strong alignment between model's rationales, probabilities and content; supporting transparent, valid inferences characterized by self-consistency and essay-specificity.

- **Exploring SHAP Value Interpretations of Machine Learning Models for English Learner Reclassification in Alaska (eBoard 14)**
  Preston Botter; Ève Ryan, University of Alaska Fairbanks

  This study investigates how SHapley Additive exPlanations (SHAP) values can enhance interpretability of machine learning models predicting reclassification of English Learners (ELs) in Alaska. By applying SHAP to Random Forest models, we examine fairness and transparency of predictors, focusing on subgroup differences and equity implications for Indigenous and non-Indigenous students.

- **AI Readiness among US Adults: Evidence from US PIAAC 2023 (eBoard 15)**
  Sneha Roy, Kansas State University; Haijun Kang, Kansas State University

  Artificial Intelligence (AI) technologies are integrated into adult lives for everyday decision-making. A large proportion of adults struggle to reap the benefits from it. With this context, this study develops an AI Readiness Index to empirically measure an adult's readiness for the adoption of AI technologies.

- **Evaluating Machine Learning Models for Item Difficulty Prediction with Feature Selection Methods (eBoard 16)**
  Juyoung Jung, University of Iowa; Yeonju Lee, University of Iowa; Ae Kyong Jung, University of Iowa; Seungwon Shin, University of Iowa; Hyung Jin Kim, National Council of State Boards of Nursing; Won-Chan Lee, University of Iowa; Sun Kim, Chungnam National University; Jae-Chun Ban, Chungnam National University

  This study investigated the ability of machine learning models to predict item difficulty for a mathematics assessment. The dependent variables were defined as CTT-based proportion correct and IRT-based difficulty parameters. Two feature selection methods were used. The support vector regression model demonstrated superior predictive accuracy on the fixed test form.

- **The Effect of Different Text Processing Techniques on Language Model Classification (eBoard 17)**
  Zhifei Li; Ariel Aloe, University of Iowa; Sarah Osaro, University of Illinois Urbana-Champaign

  This study emphasizes the importance of text preprocessing techniques in text classification models. We explore the effect of tokenization, stop-word removal, and stemming techniques on language model classification's accuracy and efficiency by using R functions.

- **Comparing SME and LLM Translations of Test Items: Content and Psychometric Evaluation (eBoard 18)**
  Oscar Rios, University of California - Davis; Tony Albano, University of California - Davis

  Accurate translation of test items is essential for fairness in multilingual assessments, yet current workflows are resource-intensive and depend heavily on bilingual translators. Advances in large language models (LLM) offer the potential to streamline translations. This study examines the quality and psychometric comparability of 50 human- and LLM-assisted translated items.

Assessment and Measurement to Support Classroom Learning (AMSCL) SIGIMIE Meeting

8:45 AM – 9:45 AM
Intercontinental Los Angeles Downtown, Floor 6: Royal

The Assessment and Measurement to Support Classroom Learning (AMSCL) SIGIMIE focuses on how classroom assessment and educational measurement interact to promote student learning.

The purposes of the SIGIMIE include initiating, reviewing, and disseminating high-quality theory, research, policy, and practice, and supporting scholars in the field, including early career scholars and graduate students. To serve these purposes, the SIGIMIE promotes rigorous research into classroom assessment practices, using all legitimate research approaches and methods, and active collaboration with other SIGIMIEs and groups outside of NCME that share an interest in promoting research into classroom assessment.

Coffee Break

Social
9:00 AM – 10:00 AM
Intercontinental Los Angeles Downtown, Floor 5 : Wilshire Grand Prefunction

Hot coffee & tea will be available.

Graduate Student eBoards: Validity, Score Reporting, and Accountability

Graduate Student Electronic Board Session
9:45 AM – 10:45 AM
Intercontinental Los Angeles Downtown, Floor 5 : Hancock Park

Presentations:

1. **Tackling Issues in Survey Validation: Leveraging Small Samples and Mixed-Format Questions**
   Kaylena Mann, Boston College; Nan Yang, Boston College; Yan Leigh, Boston College

   This study pilots a school support staff survey examining their experiences with a school-based intervention. Reliability and exploratory factor analyses provided encouraging initial evidence. Further validation through cognitive interviews and comparisons with other stakeholder groups will aid in item refinement and contribute to broader survey validation literature.

2. **Psychometric Validation and Measurement Invariance of the Bangla Grit-O Scale**
   Roti Chakraborty, Georgia State University, American Institutes for Research; Hongli Li, Georgia State University; Michael Frisby, Georgia State University

   This study validates the Bangla Grit-O scale for Bangladeshi university students through confirmatory factor analysis and measurement invariance across gender. Findings will confirm the scale's structural validity and fairness, establishing it as a culturally appropriate grit measure in Bangladesh's higher education and advancing grit's cross-cultural research in non-Western contexts.

3. **Assessing youth aggression in sport: A systematic review of validated aggression scales**
Levone Lee, University of Kentucky; Tarkington Newman, University of Kentucky; Matt Moore, University of Kentucky; Xiafei Wang, University of Kentucky

   This systematic review critically examined scales measuring youth attitudes toward aggression. The quality and validity of included articles were evaluated using the Downs and Black checklist and QUADAS. Preliminary findings indicated inconsistent conceptualization and methodological limitations. This review underscores the need for theoretically grounded instruments reflecting contemporary scale development theory.

4. **Three New Student Agency Scales: Measuring Conditions, Constraints, and Attitudes**
Nicholas Balisciano, Harvard University

   Several scales have been developed to measure the psychological facet of student agency, but not environmental conditions or agency behavior. Using mixed-methods development, survey data, and psychometric analyses (including CFA, IRT, and DIF), I refined and thoroughly analyzed three scales that will enhance how we measure student agency.

5. **Bilingual and Culturally Responsive Literacy Assessments: A Field Study in India**
Pooja Nagpal, University of Sydney

   Can bilingual, culturally grounded assessments offer fairer ways to measure foundational literacy in multilingual Indian classrooms? This randomized study compares monolingual and responsive formats using IRT modeling, DIF analysis, and cognitive interviews. Preliminary findings (January 2026) will inform more equitable assessment design across the Global South, aligned with SDG 4.1.1.

6. **Development of a Valid and Reliable Measurement for Adult Social Emotional Competency**
Angie Williams, University of Kansas

   This paper presents the development and validation of a new adult social-emotional competency scale. Using expert review, pilot testing, and advanced psychometric analyses (CTT, IRT), the tool establishes reliability, validity, and cultural relevance. Findings provide researchers and practitioners a robust measure to support well-being, education, workplace, and healthcare outcomes.

7. **Computing Comparable Scores on Culturally-Responsive Assessments Using a 2PL Model**
Anna Nasyrova, University of Massachusetts, Amherst; Scott Monroe, University of Massachusetts, Amherst

   Sinharay and Johnson (2024) proposed psychometric models for computing comparable scores for culturally-responsive assessments, to ensure fairness for examinees from all backgrounds. We extend this work by using a 2PL model, instead of a Rasch model, in the modeling framework. Key results concern the accuracy and comparability of estimated scores.

8. **Practice Makes Progress: The Impact of Ohio's Readiness Assessments on Algebra Performance**
Julie Snipes, University of Delaware

   This study investigates whether Ohio's Readiness Assessments improved Algebra I performance in one district. Using longitudinal data, growth percentiles, and regression analyses, preliminary results show positive effects. Validation with an independent audit benchmark will test whether gains reflect genuine learning rather than test-specific score inflation, providing evidence of instructional value.

9. **Validity Evidence in Personality Assessment: A systematic Review**
Tasnim Altamimi, University of North Carolina at Greensboro; Joe Sandoval, University of North Carolina at Greensboro; Talal Alzabidi, University of North Carolina Greensboro

This systematic review-in-progress evaluates validity evidence for MMPI-2, R-PAS, and Big Five inventories. Guided by the Standards framework, we examine content, response processes, internal structure, relations, and consequences. Methods include factor analysis, IRT, and DIF. Preliminary findings suggest complementary strengths, context-sensitive limitations, and clear next steps for continued screening work.

10. **A Critical Review of Fairness in Validity Frameworks for Social Justice in Testing**
Donather Magabe, University of North Carolina at Greensboro; Talal Alzabidi, University of North Carolina Greensboro

In this study, we conducted a critical review of six selected works on validity and fairness in educational testing. Our analysis examined how fairness is conceptualized and represented in major validity models, highlighted limitations in their treatment of fairness, and proposed future directions for explicitly linking fairness to validity.

11. **Assessing Pro-Social Interdisciplinary Research and Extension Education for STEM Undergraduates**
Erica Light, UMass Amherst

This quasi-experimental mixed-methods study recruited undergraduate STEM students in internships, inquiry courses, and independent studies. Measurements of STEM identity, self-efficacy, belonging, and autonomy are supplemented with narrative-based qualitative interviews and self-described identities and demographics. Analyses include structural equation modeling (quantitative) and deductive pattern coding (qualitative) guided by self-determination theory.

12. **A Nonlinear Approach for Setting ESSA Long-term Academic Achievement Goals**
Bradley Madden, University of North Carolina at Greensboro

Long-term academic goals, as required by the Every Student Succeeds Act, often rely on linear trajectories and percent of proficient students. This study proposes an alternative non-linear, asymptotic model toward a long-term goal that uses historical data and mean scale scores to set goals which are personalized to accountability groups.

13. **Beyond Traditional IRT: Integrating Summed Scores and IRT Proficiency Scores**
Qiao Liu, UNC Charlotte; Stella Kim, University of North Carolina Charlotte; Won-Chan Lee, University of Iowa

This study introduces an integrated scoring method that combines IRT proficiency scores with number-correct scores for sections or item formats that only partially meet IRT assumptions. Simulation results show the proposed method reduced error under model violations, offering a practical alternative for mixed-format tests and tests with misfitting items.

14. **Justice Oriented Systemic Review for Measurement: Belonging as Illustration**
Catherina Villafuerte, University of Connecticut

This paper introduces a justice oriented review protocol that yields measurement ready outputs. Using higher education belonging as illustration, the method produces a content blueprint, a validation and fairness inventory, and a consequences ledger aligned with the Standards. Protocol steps, reliability safeguards, and portability are reported.

AI Item Difficulty Modeling

Coordinated Paper Session
9:45 AM – 11:15 AM
Intercontinental Los Angeles Downtown, Floor 5 : Boyle Heights

We present the results of recent work on using AI to simulate student responses for the purposes of informing test development. This work seeks to estimate item difficulty parameters by calibrating new items using AI simulated responses and their scores. Our proposal includes papers on using Fine-tuned Generative Large Language Models to simulate student free-form constructed responses, an linguistic analysis of those responses, and a validation of the authenticity of the responses by educators. We also contrast this approach against classical regression-based approaches and present extensions of this work to other item types. We believe this session presents a comprehensive picture of using AI simulated student responses for difficulty modeling covering concept, implementation, and validity.

This project has been a coordinated effort between Hawaii's Department of Education, Cambium Assessment, The Smarter Balanced Consortium, and the University of Massachusetts Amherst.

Chair:
Christopher Ormerod

Presentations:

1. **Calibrated Student Simulation for Short-answer Item Difficulty Estimation**
   Andrew Lan, University of Massachusetts, Amherst; Alexander Scarlatos, University of Massachusetts Amherst; Nigel Fernandez, University of Massachusetts Amherst

2. **A Linguistic Analysis of AI Simulated Student Responses**
   Mackenzie Young; Zachary Schultz, Cambium Learning Group, Inc.; Amy Burkhardt, Cambium Assessment; Honeiah Karimi, Cambium Assessment; Suhwa Han, Cambium Assessment Inc.

3. **Educator Validation Study of AI-generated Student Responses**
   Rochelle Michel; Lynelle Morgenthaler, Smarter Balanced; Mary Cochron, Metimur Educational Measurement; Jason Varcoe, Smarter Balanced

4. **Integrating LLM-Based Predictions with Bayesian Estimation for Item Parameter Prediction**
   Suhwa Han, Cambium Asessment; Frank Rijmen, Cambium Assessment

5. **Using Generative Models to Predict Response Probabilities to Multiple Choice Questions**
   Christopher Ormerod; Suhwa Han, Cambium Assessment Inc.

AI, Machine Learning, & Natural Language Processing Research

Individual Paper Session
9:45 AM – 11:15 AM
Intercontinental Los Angeles Downtown, Floor 7: Roosevelt B

Presentations:

1. **Identifying Writing Strategies in Educational Assessments with an Unsupervised Learning Measurement Framework**
   Cheng Tang, The University of Georgia; Jiawei Xiong, Curriculum Associates; George Engelhard, University of Georgia

   This study proposes a computational framework using natural language processing and unsupervised learning to identify student writing strategies. Analyzing 406 essays, we find that text complexity and evidence use significantly predict writing proficiency. This offers a diagnostic approach to assessment beyond single scores and can inform targeted writing instruction.

2. **Investigating Learning Patterns of Students with Disability using Machine Learning Algorithms**
   Toshiko Kamei, University of Melbourne; Sadia Nawaz, Monash University; Francis Anthony, Monash University

   This paper investigates the use of Machine Learning Algorithms to analyse assessment data of students with disability in thinking skills and literacy. Findings indicate potential insights provided by using such methods to describe the learning and development of students with disability to support educational practice and policy.

3. **Automating Cognitive Domain Classification in TIMSS**
   Ummugul Bezirhan, Boston College; Matthias von Davier, Boston College

   Automated methods for classifying TIMSS items into cognitive domains using item content were evaluated using TF-IDF baselines, enhanced feature engineering, and transformer models. Enhanced TF-IDF achieved the strongest performance surpassing transformer models. Embedding analyses revealed minimal separation across cognitive domains underscoring fundamental construct level challenges.

4. **Using Convolutional Neural Networks to Detect IRT Item Misfit**
   John Donoghue, ets

   This paper examines using Convolutional Neural Networks (CNNs) to detect IRT item-level misfit. Over 90,000 item-fit plots were stored in individual jpeg files. Two-layer and three-layer CNNs were fit using RELU activation and 2 x 2 MaxPool. Test data classification accuracy of approximately 97% was obtained, with two-level models preferred.

5. **AI-Driven Predictions of Mathematics and Science Exam' Results in the UK**
   Sebastian Nastuta, Pearson Education UK

   Several supervised machine learning algorithms were used to develop a methodological framework able to predict future grades for the Mathematics and Science GCSE exam results in England. Comparing the predicted grades with the actual obtained grades, we concluded that different algorithms provide predictions with between 80% and 95% accuracy.

AI-Assisted Assessment Development

Individual Paper Session
9:45 AM – 11:15 AM
Intercontinental Los Angeles Downtown, Floor 5 : K-Town

Chair:
Henry Makinde

Discussant:
Andre Rupp (Center for Assessment)

Presentations:

1. **Efficient Detection of Bad Items with Nonlinear Scalability Coefficients**
   Michael Hardy, Stanford University; Benjamin Domingue, Stanford University

   This research introduces novel, computationally efficient nonparametric metrics for detecting flawed psychometric items. By leveraging inter-item monotonic relationships through isotonic regression, these methods demonstrate superior efficiency (AUC) in identifying bad items compared to traditional approaches, especially in complex, large-scale, or unconventional assessment scenarios like AI benchmarks.

2. **LLM-Driven Automated Item Generation for High-Stakes STEM Assessment**
   Moses Omopekunola, Center of psychometrics and measurements in education, HSE University, Moscow, Russia.

   This study evaluates LLM-AIG for high-stakes physics assessment. The generation workflow combines Bloom's taxonomy, learning objectives, and blueprint to generate multiple-choice items for UTME. The pilot study (N=527) produced 35 vetted items with strong reliability and model fit, informing scalable LLM-AIG workflows for operational testing.

3. **Comparing Expert- and AI-Based Alignment of Two Measures of Essential Skills**
   Fernando Mena, University of Massachusetts Amherst; Kate Walton, ACT

   This study compares SME and GPT-5 alignments of two essential skills measures. GPT-5 demonstrated near-perfect convergence and greater internal consistency than SMEs. Used in a human-in-the-loop framework, GPT-5 provides a scalable pre-screening tool that reduces workload and costs while maintaining expert oversight and validity in assessment alignment.

4. **Towards an automated quality-control pipeline for CPS items using LLMs**
   Yu Wang, New York University; Madhumitha Gopalakrishnan, New York University; Yoav Bergner, New York University

   State-of-the-art language models, impressive in many respects, fail even simple item-generation tasks that involve interdependence between test-takers, i.e., collaborative problem solving. Building on previous workaround strategies to overcome reasoning errors, we examine the potential of LLMs to iteratively improve CPS items with a feedback pipeline.

5. **From Creation to Evaluation: A Systematic Human-in-the-Loop Approach to Automatic Item Generation**
   Guher Gorgun, Leibniz Institute for Science and Mathematics Education; Marlit Lindner, IPN - Leibniz Institute for Science and Mathematics Education

   Our experimental study tested 4 LLMs and 10 different prompting strategies to systematically evaluate the best approaches to automatic item generation. As human involvement increasingly shifts from item creation to evaluation, we aim to advance the ground truth rating process by proposing expanded item evaluation guidelines and criteria.

Admissions and Higher Education Research

Individual Paper Session
9:45 AM – 11:15 AM
Intercontinental Los Angeles Downtown, Floor 5 : Silver Lake A

Chair:
Emily Shaw (College Board)

Discussant:
Emily Shaw (College Board)

Presentations:

1. **Impact of COVID-19 Instructional Changes on High School Grade Inflation**
   Comfort Omonkhodion, University of Central Florida; Edgar Sanchez

   This study utilized a difference-in-differences design to isolate the impact of the shift in learning modalities (associated with the pandemic) on the rate of grade inflation. Results showed that four of the learning modalities experienced larger grade inflation than schools that remained in-person during the pandemic.

2. **Rethinking Admissions Metrics: Socioeconomic Status and Standardized Testing – Meta-Analysis**
   Kiya Ma, University of Kansas

   This meta-analysis examines the relationship between SES and standardized testing. Using a random-effects model, and results show a moderate, significant correlation (effect size = .36). Findings suggest SES reflects contextual advantages rather than direct causality. The study supports holistic admissions to address SES-related disparities and calls for equity-focused policies.

3. **Using Fairness-Aware vs. Traditional Models for Predicting College Freshman GPA**
   Edgar Sanchez

   How can colleges predict FYGPA fairly after affirmative action's end? We compare logistic regression models (with and without race/ethnicity) against a fairness-aware machine learning model. Using academic and demographic data, we evaluate bias and accuracy. Results show minimal loss in accuracy without race and limited advantages to more complex models.

Automated Scoring Research: Security, Efficiency, and Interpretability

Individual Paper Session
9:45 AM – 11:15 AM
Intercontinental Los Angeles Downtown, Floor 5 : Ladera Heights

Chair:
HyunJoo Jung (Inha University)

Discussant:
Corey Palermo (Measurement Incorporated)

Presentations:

1. **Compressing Dimensions, Expanding Access: Efficient Transformer Models for Automated Essay Scoring (AES)**
   Yi Gui, Measurement Incorporated (MI)

   This study optimizes transformer-based Automated Essay Scoring by compressing DeBERTa-V3 Large embeddings via PCA and model distillation. Results show retaining 95% variance reduces embeddings to 152 dimensions with minimal accuracy loss (ΔQWK=-0.019). Distillation achieves near-native performance (QWK≈0.820) while cutting embedding generation time by 60%, enabling efficient large-scale deployment.

2. **LinguaPII: Multilingual PII Detection for Secure Automated Scoring in ILSAs**
   Ji Yoon Jung, Boston College; Ummugul Bezirhan, Boston College; Matthias von Davier, Boston College

   Protecting personally identifiable information (PII) is critical for the secure automated scoring in ILSAs. We introduce LinguaPII, a novel PII detection framework that integrates regex-based patterns with fine-tuned multilingual named entity recognition. This supports the safe implementation of multilingual automated scoring while upholding data privacy standards.

3. **Balancing Performance and Interpretability in Automated Essay Scoring: the RuleFit Algorithm**
   Xinchu Zhao, Roblox; Matt Emery, Roblox; Philip Simmons, Roblox; Sizheng Zhu, Roblox; Erica Snow, Roblox; Jack Buckley, Roblox

   This paper introduces the RuleFit algorithm to the domain of Automated Essay Scoring (AES). By combining RuleFit with both rubric-aligned and general NLP features, our model produces transparent, rule-based predictions. This approach yields a fully interpretable scoring system, with performance comparable to existing methods.

4. **Using GPT-4 for Automated Essay Scoring: Accuracy and Fairness across Student Groups**
   Yue Huang, Measurement Incorporated; Joshua Wilson, University of Delaware; Corey Palermo, Measurement Incorporated

   This study evaluates GPT-4 for automated essay scoring (AES) using prompting, fine-tuning, and fairness analyses. Prompting improved human-machine agreement, though still below feature-based models. Fine-tuned GPT-4 outperformed all other AES methods, showed fairness for special education students, and no significant bias for English learners, though human-level agreement was not reached.

Equating Research

Individual Paper Session
9:45 AM – 11:15 AM
Intercontinental Los Angeles Downtown, Floor 7: Roosevelt A

Chair:
Stella Kim (University of North Carolina Charlotte)

Discussant:
Stella Kim (University of North Carolina Charlotte)

Presentations:

1. **Evaluating the Performance of Equating Methods for Alternate Assessments**
   Daniel Adams, ETS

   Maintaining score comparability across administrations is essential for the validity of state K-12 alternate assessments. This study compares both CTT equating methods and IRT equating methods for a state K-12 alternate assessment, highlighting differences in stability and proficiency decisions when equating across test administrations.

2. **Optimizing Anchor Strategies for Pre-Equated Assessment Systems: Addressing Irrevocable Field Test Calibration**
   Soo Ingrisone, HumRRO; Kuo-Feng Chang

   Pre-equated systems enable immediate scoring but create irreversible calibration risks without validation opportunities. This simulation compares anchor strategies using robust-Z and D-square statistics across linking methods. Method-specific performance varies by ability differences and linking approaches, providing evidence-based practical guidance for minimizing calibration errors in operational banks.

3. **Item Difficulty Equating: A Comparison of Three Methods**
   Dongmei Li; Shalini Kapoor

   This study evaluates linear equating, EIRC, and IRT methods for adjusting the sample-dependent classical item difficulty statistic—p values. Using data from a large-scale assessment and a chained equating design, it examines each method's ability to produce sample-invariant p values to support test development, including parallel form construction.

4. **An Investigation of the Reused Form Method in Detecting the Test Scale Drift**
   Feifei Li, Educational Testing Service; Longjuan Liang, Educational Testing Service

   This study employed simulation to examine the effectiveness of the reused form design in detecting scale drift. Two factors were manipulated: the percentage of reused items and the percentage of drifted items. Results indicate that the ability to detect scale drift diminishes when the percentage of reused items is small.

In Memory of Robert J. Mislevy: A Legacy of Innovations in Measurement

Organized Discussion
9:45 AM – 11:15 AM
Intercontinental Los Angeles Downtown, Floor 5 : Westwood

In 2025, the field of educational measurement lost one of its most influential voices, Dr. Robert J. Mislevy, a visionary scholar whose groundbreaking contributions have shaped modern educational measurement and psychometrics. Throughout his remarkable career, Dr. Mislevy advanced psychometrics and measurement theory through a unique blend of technical rigor and conceptual insight. This organized discussion session is dedicated to commemorating Dr. Mislevy's enduring legacy. Colleagues, collaborators, and former students will come together to reflect on his profound impact on the field and to honor the ideas, mentorship, and intellectual leadership he so generously shared.

Chair:
Jiangang Hao (ETS)

Discussant:
Kadriye Ercikan (ETS Research Institute)

Presenter(s):
Kadriye Ercikan, ETS Research Institute; Rebecca Zwick, University of California, Santa Barbara; James Pellegrino, University of Illinois at Chicago; Russell Almond, Florida State University; Alina von Davier, Duolingo; Jiangang Hao, ETS; Howard Everson, City University of New York; Heather Buzick, ACT

Invited Session: Beyond "National" Measurement: An Introduction to International Assessment Research and Practices

Invited Session
9:45 AM – 11:15 AM
Intercontinental Los Angeles Downtown, Floor 7: Hollywood Ballroom II

Chair:
Susan Lyons (Lyons Assessment Consulting)

Presentations:

1. **A History of the International Test Commission**
   April Zenisky, University of Massachusetts, Amherst

2. **Laws are National, But Guidelines are Not: Moving toward International Standards**
   Kurt Geisinger

3. **The Role of the New PISA Quality Standards in Promoting Fairness**
   Javier Suárez-Álvarez, University of Massachusetts, Amherst

4. **Assessing Languages Across Borders: Validity Issues in Multimodal Representations of Language Proficiency**
   Angel Arias, Carleton University

5. **The International Test Commission Guidelines: Promoting Fairness in Assessment Since 1993**
   Stephen Sireci, University of Massachusetts Amherst

Scoring Methodologies and Best Practices for Score Reporting

Individual Paper Session
9:45 AM – 11:15 AM
Intercontinental Los Angeles Downtown, Floor 5 : Silver Lake B

Chair:
Marisol Kevelson (ETS)

Discussant:
Tim Moses (Buros Center for Testing)

Presentations:

1. **A Simple Augmented Subscore Method for Computerized Adaptive Tests**
   Michael Peabody, IXL Learning

   Reporting subscores to users can be challenging and existing methods for augmenting subscores rely on group-level statistics. This study proposes a subscore augmentation method that utilizes only information from an individual's testing event.

2. **Evaluating Best Practice Criteria for Concordance Studies**
   Anastasia Ulicheva, Pearson; Sarah Hughes

   The present study evaluates best-practice criteria for test score concordance. Reanalysis of a large dataset examines sample size, counterbalancing, retest intervals, and population invariance. Results show that stability varies across score regions, refining methodological standards for concordance studies that inform high-stakes educational and immigration decisions.

3. **Which Scoring Methods Should be Used for Shortened Scales?**
   Wenshuo Li, McGill University; Okan Bulut, University of Alberta

   This study examines the properties of four existing scoring methods for short-form scales: factor scores estimated from confirmatory factor analysis, optimal linear combination scores, supervised machine learning-predicted scores, and summed scores. Results are evaluated regarding consistency with original scales, reliability, and abilities to retain correlational patterns under varying conditions.

4. **Beyond Binary Categories: A Comprehensive and Fair Approach to Universal Reading Screening**
   Julian Siebert, University of California, San Francisco; Mónica Zegers, UCSF; Marilu Gorno Tempini, University of California, San Francisco

   We propose and illustrate a comprehensive framework for the development of universal (reading) screeners in educational settings. This approach posits that---in order to yield fair and locally actionable information---screener development should be guided not only by the maximization of classification accuracy, but also by instructional and clinical relevance.

5. **Reliability and Accuracy in Standard-Level Mastery Classification: Subscores vs. Item Maps**
   Garron Gianopulos, Cambium Assessment; Christina Schneider, Cambium Assessment, Inc.

   Two score reports were compared for standard-level mastery decisions: subscores and item maps. Subscores lacked reliability ($r < .60$), but classification decisions based on item maps showed high accuracy (>.87), supporting their use for student-level decisions. The limitations and implications of this study will be discussed.

The Content and Psychometric Partnership: A Panel Discussion (Invited Session)

Invited Session
9:45 AM – 11:15 AM
Intercontinental Los Angeles Downtown, Floor 6: Majestic

Join content developer professionals and psychometricians for a forward-looking conversation about the partnership at the core of assessment and how can we strengthen this collaboration in ways that enhance validity, fairness, and defensibility. This panel will explore the evolving professional identity of content developer specialists, the potential for shared competencies and standards, and how professional organizations such as NCME might better support this community. Panelists will also consider how technology is influencing the craft of content development and what skills and structures will be essential in the years ahead.

Chair:

Amy Hendrickson (The College Board)

Panelists:
Susan Davis-Becker, ACS Ventures, LLC
Katie Schmidt, College Board
Mary Veazey
Catherine Welch, University of Iowa
Marjorie Wine, Accessible Teaching, Learning and Assessment Systems (ATLAS), University of Kansas

The Revision of the Testing Standards: Applications (Invited Joint AERA/NCME Session)

Organized Discussion
9:45 AM – 11:15 AM
Intercontinental Los Angeles Downtown, Floor 5 : Wilshire Grand Ballroom III

This is the third of three joint sessions designed to update the memberships of AERA and NCME and invite input on the revision of the Standards for Educational and Psychological Testing. This session will focus on the applications chapters, with particular attention to psychological testing and assessment, workplace testing and credentialing, educational testing and assessment, and use of tests for program evaluation, policy studies and accountability. The Joint Committee (JC) has been working on revising the 2014 Standards. In this session, participants will receive an update on the revision process and learn about major changes proposed for the application chapters. Most importantly, the session will provide time to hear directly from participants—their reactions to the updates as well as their questions, hopes, and concerns about the ongoing revision.

Chair:
Ye Tong (National Board of Medical Examiners)

Discussant:
Andy de los Reyes

Presentations:

1. **The Revision of the Testing Standards: Applications (Invited Joint AERA/NCME Session)**
   Ye Tong, National Board of Medical Examiners; Frank Worrell; Nathan Kuncel; Ellen Forte, edCount, LLC; Laura Hamilton, National Center for the Improvement of Educational Assessment, Inc.; Andy de los Reyes; Kristen Huff, Curriculum Associates

   This is the third of three joint sessions designed to update the memberships of AERA and NCME and invite input on the revision of the Standards for Educational and Psychological Testing. This session will focus on the applications chapters, with particular attention to psychological testing and assessment, workplace testing and credentialing, educational testing and assessment, and use of tests for program evaluation, policy studies and accountability.

   The Joint Committee (JC) has been working on revising the 2014 Standards. In this session, participants will receive an update on the revision process and learn about major changes proposed for the application chapters. Most importantly, the session will provide time to hear directly from participants—their reactions to the updates as well as their questions, hopes, and concerns about the ongoing revision.

   Chair:
   Tong, Ye - National Board of Medical Examiners

   Discussant:
   de los Reyes, Andy

AI-Driven Interactive Speaking and Math Assessments: Innovations in Design and Scoring

Coordinated Paper Session
11:30 AM – 12:45 PM
Intercontinental Los Angeles Downtown, Floor 5 : Ladera Heights

This coordinated session highlights advances in AI-driven interactive assessment at Duolingo and Khan Academy, showcasing innovations in adaptive task design, automated scoring, and the use of large language models (LLMs) to capture deeper aspects of student performance.

The first presentation introduces Duolingo's interactive speaking assessment, an AI-managed system where a character engages test takers in simulated interviews. It generates authentic topics, adaptive prompts, and expected response elements to evaluate task achievement. Validation results showed AI-human agreement comparable to human-human agreement, supporting AI's potential to deliver valid, adaptive speaking assessments that mirror human interviewers.

The second, third, and fourth presentations focus on Khan Academy's "Explain Your Thinking" (EYT) items, which use AI chatbots to engage students in multi-turn dialogues and probe conceptual reasoning in mathematics. LLMs score students' understanding based on these exchanges. The second study examines rater–LLM agreement, the third explores strategies to optimize LLM scoring (e.g., sampling, tuning parameters), and the fourth investigates methods for estimating confidence levels in LLM scores to improve reliability.

Together, these studies demonstrate how AI can enhance the reliability, validity, and design of interactive assessments, enabling scalable and authentic evaluation of complex skills across domains.

Chair:
Jing Chen

Discussant:
Kristen Dicerbo (Khan Academy)

Presentations:

1. **Validity of On-the-Fly Automated Scoring for Adaptive Interactive Speaking Assessments**
   Yigal Attali, Duolingo

2. **Explain Your Thinking: Using AI Conversations to Assess Mathematical Understanding**
   Jing Chen, Khan Academy

3. **Does LLM Self-Consistency Improve Automated Short Answer Grading?**
   Scott Frohn, Khan Academy

4. **Model Confidence as a Routing Signal: Optimizing Automated Scoring Through Selective Triage**
   Tyler Burleigh, Khan Academy

Advancing Digital Skills Assessment: Integrating Tests, Performance Data, and Process-Oriented Analyses

Coordinated Paper Session
11:30 AM – 12:45 PM
Intercontinental Los Angeles Downtown, Floor 5 : Silver Lake A

As higher education and vocational education and training (VET) systems undergo rapid digital transformation, the valid assessment of students' skills remains a pressing challenge. Traditional measures often fail to capture the complex, situated, and evolving nature of competencies required for academic success and workforce readiness.

This symposium presents innovative approaches from three international research collaboratives for validly assessing students' digital skills across diverse contexts. Contributions span from entry diagnostics using self-reports and standardized tests of first-year students, to process-oriented performance assessments with eye-tracking during real web searches and qualitative analyses of think aloud protocols. Moreover, the symposium offers an international outlook by introducing the OECD's PISA VET initiative, which develops novel frameworks and comparative assessments for vocational competencies worldwide.

By integrating tests, behavioral data, and process-based evidence, the symposium highlights how methodological innovations can generate more valid and reliable measures of digital and vocational skills. Discussions will address challenges of assessment design, cross-cohort comparability, and the implications of AI integration for student learning and equity. Together, the papers illustrate both the opportunities and challenges of current assessment approaches, and they underscore the importance of international collaboration and interdisciplinary perspectives for advancing skill measurement in rapidly changing educational environments.

Chair:
Olga Zlatkin-Troitschanskaia

Discussant:
Patricia Alexander (University of Maryland)

Presentations:

1. **Assessing key study preconditions in the digital age – Cohort comparison of COR&AI-skills**
   Jasmin Reichert-Schlax, Johannes Gutenberg University Mainz; Olga Zlatkin-Troitschanskaia, Johannes Gutenberg University Mainz; Lukas Trierweiler, Johannes Gutenberg University Mainz; Marie Nagel; Lisa Martin de los Santos Kleinz, Johannes Gutenberg University Mainz

2. **Modeling Eye-Tracking Profiles to Identify Students' Processing Strategies in Online Reasoning Tasks**
   Andreas Maur; Anika Kohmer, Johannes Gutenberg University Mainz; Maruschka Weber, Goethe University Frankfurt; Verena Klose, Goethe University Frankfurt; Stefan Küchemann, Ludwig-Maximilians-Universität München; Ann-Kathrin Kunz, Johannes Gutenberg University Mainz; Verena Ruf, Ludwig-Maximilians-Universität München; Yavuz Dinc, Ludwig-Maximilians-Universität München

3. **University Students' Recognizing and Considering Consequences in Critical-Thinking Essays with Think-Alouds**
   Katharina Depré, Johannes Gutenberg University Mainz; Olga Zlatkin-Troitschanskaia, Johannes Gutenberg University Mainz; Dominik Braunheim, Johannes Gutenberg University Mainz; Thiemo Hagen, Johannes Gutenberg University Mainz; Richard Shavelson, Johannes Gutenberg University Mainz

4. **Reimagining Skills Assessment for Career and Technical Education: An Example from PISA-VET**
   Ou Lydia Liu, ETS; Jonathan Rochkind, ETS

An NCME Debate Session: The AI Landscape for Industry and Academia (Featured Session Format)

Invited Session
11:30 AM – 12:45 PM
Intercontinental Los Angeles Downtown, Floor 5 : Wilshire Grand Ballroom III

This 75 minute NCME Debate session will be both educational and entertaining! There will be two 25-minute debates between experts in the field, each one covering a different aspect of the AI landscape: 1) Impact of AI on Practice/ Industry and 2) Impact of AI on Academia. Audience participation will be encouraged, and there will be time for Q&A at the end.

Chair:
James Wollack (University of Wisconsin - Madison)

Debaters:
Joshua Goodman, National Commission on Certification of Physician Assistants
Ummugul Bezirhan, Boston College
Laine Bradshaw, Pearson
Hong Jiao, University of Maryland

Applications of Psychometric Methods in Digital Learning Systems

Coordinated Paper Session
11:30 AM – 12:45 PM
Intercontinental Los Angeles Downtown, Floor 5 : Silver Lake B

Placing students into digital lessons is a common digital assessment use case. These digital lessons offer rich opportunities for psychometric application. This session highlights applications of psychometrics with digital lessons to draw inferences about student learning, and the assessment and instruction systems themselves.

Two papers leverage data from Curriculum Associates' i-Ready assessment and instruction system. One paper uses logistic regression to link lesson difficulties to the assessment scale. Results reveal nuanced patterns in difficulty across the lesson sequence and offer insights informing student placement and lesson sequencing. The second paper applies item analyses to items in digital lessons. We discuss the impact of instruction on item statistics and propose approaches for flagging problematic items that consider the unique digital instruction context.

Papers three and four leverage data from Edmentum's Exact Path personalized learning platform. In paper 3, two methods of using Exact Path assessment scores to determine student placements are compared in terms of the impact on subsequent student outcomes. This paper provides additional insight into whether acceleration or remediation is more beneficial for students of various initial achievement levels. The final paper explores student outcomes when customers switch placement methods from an interim assessment to a state through-year assessment.

Chair:
Logan Rome

Discussant:
Okan Bulut (University of Alberta)

Presentations:

1. **Determining Lesson Path Starting Points Using a Time-Moderated Rasch Model**
   Fusun Sahin, Curriculum Associates

2. **Item Bank Review and Monitoring in a Digital Learning System**
   Jiawei Xiong, Curriculum Associates

3. **Comparing student growth for two personalized learning placement methods**
   Benjamin Andrews

4. **Using state assessments and interim assessments to drive personalized learning**
   Sonya Powers, Edmentum

Communicating Technical Results and Concepts to Public Audiences

Coordinated Paper Session
11:30 AM – 12:45 PM
Intercontinental Los Angeles Downtown, Floor 5 : Boyle Heights

As a community of measurement experts, we are often charged with translating our assessment results or technical findings into digestible sound bites, clear visuals, or relatable analogies. This coordinated session features four efforts to communicate critical results to various public audiences from parents to policymakers. Each of the following cases provides lessons learned on turning technical concepts or results into clear, actionable information for audiences typically less steeped in psychometrics, who are balancing many, often competing, priorities:

1. Helping parents navigate school choice decisions
2. Guiding education leaders in their choice of stabilization method to improve the precision of small group estimates,
3. Revealing nuanced findings about COVID-19 pandemic academic recovery to states and their constituents, and
4. Remodeling the Nation's Report Card web reporting pages and data tools for state leaders, policymakers, and the public press to improve educational access for all students.

Each presentation describes the context, engagement with the intended audience, and factors that were found effective (or not) in translating research into practical decisions. Following the presentations, discussant Priya Kannan will highlight common themes and share insights from her reporting experience. The session will then open for audience questions and encourage attendees to share their experiences regarding effective communication.

Chair:
Katherine Castellano

Discussant:
April Zenisky (University of Massachusetts, Amherst)

Presentations:

1. **A Framework for Informing School Choice**
   Andrew McEachin, ETS Research Institute; Laura Hamilton, National Center for the Improvement of Educational Assessment, Inc.

2. **Quantifying, Reducing, and Explaining Uncertainty in Aggregate Test Score Metrics**
   Benjamin Shear, University of Colorado Boulder

3. **Using Data Visualization to Communicate Widening Achievement Disparities**
   Katherine Furgol Castellano, ETS Research Institute; Marisol Kevelson, ETS; Emily Kerzabi, ETS

4. **Contextualizing Score Interpretations on the Nation's Report Card**
   Robert Finnegan, ETS

Graduate Student eBoards: DIF, Dimensionality, Growth Modeling, and Equating

Graduate Student Electronic Board Session
11:30 AM – 12:45 PM
Intercontinental Los Angeles Downtown, Floor 5 : Hancock Park

Presentations:

1. **The Foundational Step in Equitable Assessment: Establishing Measurement Invariance Before Score Interpretation**
   Moses Mohamed, University of South Florida; Robert Dedrick, University of South Florida; Eunsook Kim, University of South Florida; Courtney Kirby, University of South Florida; Brenda Gutu, Gallaudet University

   This empirical study established measurement invariance (MI) for an attitude scale across gender through multi-group confirmatory factor analysis (MG-CFA). Results supported full scalar invariance, confirming consistent measurement properties. This validates the significant difference in latent attitude means between boys and girls, ensuring fair and meaningful group comparisons.

2. **Examining DIF Over Time using Tree Models: An Illustration and Simulation Study**
   Theode Niyirinda, University of Alabama; Stefanie Wind, The University of Alabama; Sarah McKellar, University of Alabama

   We explored the use of Rasch trees for identifying differential item functioning (DIF) over time in longitudinal survey studies. We used real data from a longitudinal study on math anxiety and a simulation study. Results support using time as a covariate in tree models for identifying time-related DIF.

3. **Clustering Non-Ordinal Covariates in Rasch Trees: Implications for DIF Detection**
   Andrew Krist; Jujia Li, University of Alabama

   This study evaluates how clustering non-ordinal covariates prior to Rasch Tree analysis influences differential item functioning detection. Using the Generic Conspiracist Beliefs Scale, we compare baseline and clustered Partial Credit Trees across bootstrap replications. Results highlight trade-offs between interpretability, stability, and DIF outcomes, giving methodological insights for practical use.

4. **Detecting SES-Related DIF in PISA 2022 Mathematics Across Countries**
   Archangel Gundula, University of Nebraska-Lincoln (UNL); Jordan Wheeler, University of Nebraska - Lincoln

   This study examines measurement invariance and differential item functioning (DIF) in PISA 2022 mathematics items across socioeconomic groups. Using multi-group confirmatory factor analysis and IRT-based DIF detection methods on data from 10 countries, it identifies SES-related bias to improve fairness and validity in cross-national educational assessments, informing equitable test design.

5. **Principals' Perception of Creativity: An Exploratory Analysis of PISA 2022**
   Daniel Oyeniran, The University of Alabama; Christoper Ocheni, The University of Alabama, Tuscaloosa; Justice Dadzie, The University of Alabama; Yusuf Isah, The University of Alabama

   This study evaluated principals' perception of creativity using an exploratory Rasch model analysis on PISA 2022 data. The Principal Creativity Perception Scale showed a two-factor structure-teachers' and students' creativity, with acceptable and consistent fit statistics across the entire population. Students' creativity levels in religious and private schools were relatively higher.

6. **Comparing Anchor Methods for Polytomous DIF Detection Using the Mantel Chi-Square Statistic**
   Leyna Kataoka, University of California, Davis; Tony Albano, University of California - Davis

   This study compares the performance of anchor methods using the partial credit model with likelihood ratio testing for polytomous DIF detection. Monte Carlo simulations are used to evaluate Type I error rates and power rates of anchor methods across various conditions.

7. **Small-Sample Ordinal EFA: Interactions Among Correlation Type, Regularization, and Extraction**
Yi-Chen Lu; Jyun-Hong Chen, National Cheng Kung University

Using high-ecological MIRT–GRM simulations of three-factor ordinal scales (N≤200), we evaluate how correlation type, shrinkage interact with extraction to affect structure and parameter recovery. With Target/Procrustes, outcomes include Heywood rates, Tucker's φ, pattern accuracy. PAF/WLS provide greater stability; Optimize item quality to improve structure recovery.

8. **Comparing EFA and NCD in Psychometric Clustering: A Machine Learning Approach**
Jingyang Li, University of Georgia; Laura Lu, University of Georgia; Yizhu Gao, University of Georgia; Pengsheng Ji, University of Georgia

This study compares Exploratory Factor Analysis and Network Community Detection in clustering performances via machine learning approaches, such as K-Means, Random Forests and Support Vector Machine. Through simulation studies, we showed that the differences between EFA and NCD not only depend on their mechanism but also on the pre-processing procedures and classifier choices.

9. **Informative priors on MIMIC model with small sample sizes and outliers**
Nancy Alila, Department of Educational Psychology, University of Georgia, Athens, GA, USA; Laura Lu, University of Georgia

This study investigates Bayesian MIMIC models under small samples and latent outliers. We examine how informative priors influence estimation, extending a previous research by introducing contamination. Results show correctly specified priors on regression coefficients reduce bias and improve robustness, highlighting the critical role of prior calibration in practice.

10. **Exploring Covariate Imbalance in Measurement Invariance: A Simulation Study**
Autumn Wild, James Madison University; Joseph Kush, IXL Learning

This simulation study investigates how sample size imbalance, factor loading magnitude, and misfit severity affect the detection of metric invariance in a one-factor, ten-indicator model using multi-group CFA. Results will inform best practices for evaluating measurement equivalence under realistic, unbalanced data conditions.

11. **Development of Mental State Terms in Elementary Students' Writing**
XINYI WANG; Hong Jiao, University of Maryland; Fang Luo, Beijing Normal University

This longitudinal study examined mental state terms (MST) usage in elementary school children' writing across six grades. Findings showed nonlinear growth patterns in overall MST. Category-specific MST analyses showed individual differences existed primarily in baseline levels rather than growth rates. The results have implications for curriculum design and early intervention.

12. **Indirect Effect Estimation with Error-Correction Models in Multilevel Mediation Analysis**
Seongmin Park, Korea University

This study employs Monte Carlo simulation to compare frequentist and Bayesian 1-1-1 multilevel mediation models for estimating the between-level indirect effect, assessing whether simpler frequentist alternatives yield less precise estimates than Bayesian approaches with default or inaccurate priors, with the aim of providing methodological guidance for applied researchers.

13. **Spatial Methods for Understanding Academic Proficiency from an Ecological Perspective**
Laura Pires Gifford, Washington State University; Brian French, Washington State University

Heterogeneity exists in academic performance across the geographic locations of schools. This variability was modeled using an innovative spatial analytic approach. Predictors of academic achievement, including health opportunity, showed different relationships in linear and geographically weighted regressions, highlighting variability across contexts and the need for these models.

14. **A Comparison of Linear and Equipercentile Equating Methods Using the Random Groups Design**
Godwin Sabboh, University of North Carolina - Greensboro

This study compares linear and equipercentile equating methods using random groups design through simulation with varying sample sizes and test lengths under the 3-PL IRT model. Results suggest linear equating should performs well under normal distributions, whereas equipercentile equating should offer greater accuracy under skewed conditions, with practical recommendations provided

15. **Psychological Safeguarding and Conversational Confidence: New Measures Predicting Parents' Politicized School Censorship**
Sarah Hatch, Harvard University; Hannah Castner, Harvard University

Significant research is needed to better understand why parents oppose and censor certain educational programming. Using Classical Test Theory and Item Response Theory, the current study develops validity evidence for two scales of parents' beliefs—Psychological Safeguarding and Conversational Confidence—which may predict parents' politicized censorship behaviors.

Unpacking Item Difficulty Predictor Relationships Across and Within Grades

Coordinated Paper Session
11:30 AM – 12:45 PM
Intercontinental Los Angeles Downtown, Floor 7: Roosevelt A

This coordinated paper session explores how item difficulty modeling (IDM) can be used to better understand the nuanced relationships among item features (predictors) and item difficulty. The underlying assumption is that these features reflect facets of a cognitive model and are related to what makes items relatively easy or challenging. Unique to this session will be that each paper includes performance level descriptors as a feature in the IDM model. Contrasting difficulty prediction models across and within grade levels will be another area of focus. This session brings together four studies that apply IDM techniques such as OLS regression, random forest models, and AI-based methods to investigate how item features function differently as predictors when examined across grades versus within grade.

Chair:
Christina Schneider

Discussant:
Jeffrey Steedle

Presentations:

1. **Item Difficulty Prediction in K–8 Assessment Using LLMs and Tree-Based Regression**
Mihye Choi, Curriculum Associates

2. **Evaluating the Functioning of Range PLDs as a Component of Cognitive Complexity**
Christina Schneider, Cambium Assessment; Honeiah Karimi, Cambium Assessment; Christy Kulczycki, Cambium Assessment, Inc.; Wei Tao, Cambium Assessment

3. **Evaluating the Functioning of PLDs for PAD-generated NGSS Science Assessments**
Melissa Fincher, edCount, LLC; Zachary Warner, New York State Education Department; Daisy Wise Rutstein, edCount, LLC; Megan Kinmartin, New York State Education Department

4. **Investigating Theory-Based Item Features for Reading: An IDM Study**
Christina Schneider, Cambium Assessment; Sangdon Lim, Cambium Assessment; Honeiah Karimi, Cambium Assessment

Validation of Teaching Performance Assessments: Simulation Based Measures of Core Teaching Practices

Coordinated Paper Session
11:30 AM – 12:45 PM
Intercontinental Los Angeles Downtown, Floor 5 : K-Town

The improvement of teaching relies on being able to define and measure it at scale, across contexts. A recent approach advanced by scholars defines teaching into core practices (Grossman et al, 2009), also known as high leverage practices (Ball & Forzani, 2011; CEC, 2024). While early research on performance assessments of these core practices is promising, there has been little attention to foundational validity issues. The field seems to be walking the same path walked nearly 30 years ago when student performance assessments came back into favor as a solution to the need for more "valid" assessment (Messick, 1994). To respond to Messick's call for appropriate assessment rationales and address yet undone validation work on core practice assessments, this session brings together four groups of researchers creating validity evidence for core teaching practices. Drawing on Kane's (2006) articulation of validity arguments, collectively, the papers address issues of construct definition, construct under-representation, construct irrelevant variance, and the scoring inference. The papers highlight the affordances and constraints of these formative performance assessments and underscore the need for thoughtful validation of all types of assessments.

Chair:
Courtney Bell

Discussant:
Scott Marion (Center for Assessment) and Jamie Mikeska (ETS)

Presentations:

1. **Defining the Core Practice of Co-Planning for Student Success**
   Jon Nordmeyer, University of Wisconsin-Madison

2. **Toward Simulation-Based Measures of Co-Planning for Student Success**
   Courtney Bell, University of Wisconsin - Madison; Jon Nordmeyer, University of Wisconsin-Madison; Mariana Castro, University of Wisconsin - Madison

3. **Validation Evidence of High-Leverage Teaching Practices in Mathematics**
   Sarah Lent, University of Wisconsin - Madison; Jessica Tierney, University of Wisconsin - Madison

4. **Measuring Core Teaching Practices: Validity Evidence from Bilingual Science Teacher Preparation**
   Mariana Castro, University of Wisconsin - Madison; Mark Olson, University of Wisconsin - Madison

Advancing Anti-Racist and Culturally Responsive Assessments Across Contexts

Coordinated Paper Session
11:30 AM – 1:00 PM
Intercontinental Los Angeles Downtown, Floor 6: Majestic

This session convenes three projects that tackle a key challenge for the NCME community: designing, evaluating, and implementing assessments that are valid, fair, and anti-racist. While interest in culturally responsive assessment has grown, questions remain about credible evidence, how to center those most impacted by testing, and how to apply frameworks at scale. Our session addresses these issues with conceptual, methodological, and practical insights that move toward solutions.

The first paper reviews U.S. medical education interventions aimed at mitigating racism in physician–patient interactions. It maps how justice-oriented frameworks define outcomes, the measures used, and the strength of claims, offering a foundation for equity-aligned evaluations.

The second paper amplifies rights-holders' voices—students, parents, and educators—through think-alouds and interviews on culturally responsive 8th-grade math items. Findings show how relevance, representation, and authenticity shape fairness and engagement, while warning against tokenism. Recommendations guide developers in scaling culturally responsive content with rigor.

The third paper documents a partnership with Chicago Public Schools to revise math assessments using the Culturally Conscious Assessment framework. It details tensions and solutions in balancing standards, cognitive demand, and cultural responsiveness in a large-scale system.

Collectively, these studies contribute strategies and frameworks for designing assessments that advance justice in education.

Chair:
Sandra Cruz

Discussant:
Susan Lyons (Lyons Assessment Consulting)

Presentations:

1. **Interrogating the Metrics: A Systematic Review of Evaluation Practices in Anti-Racism Medical Education Interventions**
   Eduardo J Crespo Cruz, University of Massachusetts, Amherst

2. **Rights-Holders' Reflections on Culturally Responsive Math Assessment Content**
   Sandra Cruz, George Washington University

3. **From Framework to Practice: A Collaborative Approach to Culturally Responsive Math Assessments**
   Jennifer Rodriguez, University of Michigan; Juan Arvizu-Sevilla, University of Illinois at Chicago

Assessing Foundational Competencies: Advancing Educational Measurement through Feedback, Alignment, and Certification

Coordinated Paper Session
11:30 AM – 1:00 PM
Intercontinental Los Angeles Downtown, Floor 5 : Westwood

In 2024, an NCME Task Force on Foundational Competencies in Educational Measurement (FCEM) published three competency domains and five competency subdomains necessary for future learning in the field (Ackerman et al., 2024; Ho et al., 2024). In this symposium, we ask whether and how FCEM domains can support assessment for various purposes, including feedback, curricular alignment, and certification. Derek Briggs, who appointed the Task Force, will lead with a presentation identifying how students and instructors can use Artificial Intelligence to assess and develop their expertise in each FCEM domain.

Anastasiya Lipnevich will discuss extensions of her research on instructional feedback with implications for assessment of foundational competencies in courses and in the workplace. Susan Davis-Becker will discuss the results of a practice analysis and its implications for FCEMs and certification of measurement professionals. And Brian Leventhal will describe how he used FCEMs to evaluate and improve alignment through the curricular and co-curricular sequence in a graduate program in educational and psychological measurement. Wenchao Ma, a co-editor of a new journal about the teaching of educational research methods, will offer his insights as a discussant.

Chair:
Andrew Ho (Harvard University)

Discussant:
Wenchao Ma (University of Minnesota)

Presentations:

1. **Using Artificial Intelligence to Support Foundational Competencies when Teaching Educational Measurement**
   Derek Briggs, University of Colorado - Boulder

2. **Shaping Competence Through Feedback: Insights for Educational Measurement and Professional Practice**
   Anastasiya Lipnevich, Queens College and the Graduate Center, City University of New York

3. **Certifying Measurement Professionals: Design of a Practice Analysis**
   Susan Davis-Becker, ACS Ventures, LLC

4. **Investigating How a Graduate Program Aligns with Foundational Competencies: Methods and Implications**
   Brian Leventhal, James Madison University

Navigating NCME Journals: Pathways to Publication (Invited Session: Publications Committee)

Organized Discussion
11:30 AM – 1:00 PM
Intercontinental Los Angeles Downtown, Floor 7: Hollywood Ballroom II

This organized panel discussion brings together editors from NCME's three flagship journals - Journal of Educational Measurement (JEM), Educational Measurement: Issues and Practice (EM:IP), and Chinese/English Journal of Educational Measurement and Evaluation (CEJEME) - to demystify the publication process. Although primarily aimed at graduate students, early-career professionals, and new NCME members, the session will help all attendees gain a better understanding of journal scope, submission strategies, and review processes that part of the publication process. Through an interactive format combining brief journal introductions with structured Q&A, the editors will provide insider perspectives on manuscript selection, common pitfalls, and emerging publication trends. The discussion will explore how each journal serves distinct yet complementary roles within the educational measurement community: JEM's focus on theoretical advances and generalizable research, EM:IP's emphasis on practical applications and policy implications, and CEJEME's bridge between international scholarship and methodological innovation. Attendees will gain practical insights into matching manuscripts to appropriate outlets, navigating peer review, and understanding editorial decision-making. This session directly supports NCME's mission of advancing measurement science by fostering the next generation of contributors while promoting fuller utilization of the organization's diverse publication portfolio.

Chair:
Christopher Runyon (National Board of Medical Examiners)

Discussant:
Jinghua Liu (The National Board of Osteopathic Medical Examiners)

Presenter(s):
Christopher Runyon, National Board of Medical Examiners; Heather Buzick, ACT; Won-Chan Lee, University of Iowa; Yi Zheng, Arizona State University; Hong Jiao, University of Maryland

Development, Validation, and Generalizability Studies of the Compass Classroom Observation Tool

Coordinated Paper Session
1:45 PM – 2:45 PM
Intercontinental Los Angeles Downtown, Floor 5 : Ladera Heights

The Compass was developed to address a documented gap in social and emotional learning measurement by providing an observational assessment that identifies effective SEL teaching and guides SEL implementation to improve elementary school students' success. Conceptualized to be a free and useful tool for schools, the Compass was developed over seven years and was informed by twenty-three school leaders, and the systematic observation of seventy-seven classroom teachers across twenty-six schools in the US. Structured observations were facilitated by Swivl technology and live observations by leaders, and recordings were iteratively scored over the project period to develop, refine, validate, and assess the generalizability of the Compass indicators on student academic, social and behavioral outcomes during elementary school. Extant measures of the classroom (CLASS), teachers (SEL beliefs, practices, demographics, and professional experiences), and students (standardized achievement tests, SDQ, and demographics) were concurrently analyzed. The resulting Compass consists of twenty-three indicators that reflect modeling, practice promotion, transfer promotion, validation, invalidation, and elaboration, demonstrating strong evidence of ecological validity and reliability in diverse school contexts. This session features three papers that document the (1) development studies, (2) psychometric validation study, and (3) generalizability study of the Compass in US elementary classrooms.

Discussant:
Michael Strambler (Yale University)

Presentations:

1. **Development of the Compass Checklist for Elementary Schools: A Teaching Tool to Scaffold Students' Social Emotional Development**
   Joanna Meyer, The Consultation Center, Yale University; Almut Zieher, Education Collaboratory at Yale University; Craig Bailey, Yale University; Cheyeon Ha, Yale University; Michael Strambler, Yale University; Christina Cipriano, Yale University

2. **Validation of the Compass Checklist: An Educator Observation Tool for Social and Emotional Pedagogy in Elementary Schools**
   Almut Zieher, Education Collaboratory at Yale University; Craig Bailey, Yale University; Sophie Barnes, Yale University; Joanna Meyer, The Consultation Center, Yale University; Cheyeon Ha, Yale University; Michael Strambler, Yale University; Christina Cipriano, Yale University

3. **Using the Compass Checklist to Support Social and Emotional Pedagogy: Implications Based on Generalizability Theory**
   Sophie Barnes, Yale University; Almut Zieher, Education Collaboratory at Yale University; Craig Bailey, Yale University; Michael Strambler, Yale University; Christina Cipriano, Yale University

Illinois's Unified Standard Setting: Innovative Strategies to Support Coherence

Coordinated Paper Session
1:45 PM – 2:45 PM
Intercontinental Los Angeles Downtown, Floor 6: Majestic

In July 2025, the Illinois State Board of Education (ISBE), in collaboration with Pearson, ACT, and the Center for Assessment, implemented a unified standard setting approach for the Illinois Assessment of Readiness (IAR), Illinois Science Assessment (ISA), and ACT. This landmark effort aligned Grades 3–11 expectations across ELA, mathematics, and science, defining four performance levels and creating new sets of Performance Level Descriptors (PLDs) including new innovative 'Samples for Success.'

This session presents three coordinated papers:

Paper 1, Building the Foundation – Fostering Coherence through Unified PLD's, Transparency, and Communication, details the creation of PLDs and the strategic communications that fostered statewide understanding and buy-in.

Paper 2, Operationalizing Coherence – Strategic Design and Methodology for Unified Standard Setting, examines the technical and logistical coordination across assessments, including innovative judgment methods, student profiles, and alignment strategies that ensured consistency from elementary to high school.

Paper 3, Empirical Evidence in Action: ACT Extensions for Unified Standard Setting Strategy, explores the integration of empirical data—benchmarks, postsecondary success probabilities, and PreACT performance—in the standard setting process for high school.

Discussant Will Lorie, a TAC member who observed the standard setting, will highlight national implications for unified standard setting practice.

Chair:
Andre Rupp (Center for Assessment)

Discussant:
Will Lorie (Center for Assessment)

Presentations:

1. **Building the Foundation – Fostering Coherence Through Unified PLD's, Transparency, and Communication**
   Will Lorie, Center for Assessment; Andre Rupp, Center for Assessment

2. **Operationalizing Coherence – Strategic Design and Methodology For Unified Standard Setting**
   Tracy Gardner, Pearson

3. **Empirical Evidence In Action: ACT Extensions For Unified Standard Setting Strategy**
   Shalini Kapoor; Joann Moore, ACT

Individual eBoards: Scaling, Equating, Linking, and High-Stakes Assessment

Electronic Board Session
1:45 PM – 2:45 PM
Intercontinental Los Angeles Downtown, Floor 5 : Hancock Park

Presentations:

1. **Grade Inflation and Academic Mismatch in College Readiness Indicators**
   Edgar Sanchez

   This study examines the growing disconnect between HSGPA and standardized test scores. Using national data from 2013–2023, we explore the prevalence of academic mismatch, its relationship to grade inflation, and demographic patterns. Results show that grade inflation contributes to mismatches, with implications for college readiness and admissions policy.

2. **Advancing Item Writing Methods with Large Language Models and Subject Matter Experts**
   Oscar Rios, University of California - Davis; Tony Albano, University of California - Davis

   In this paper, we investigate generative large language models (LLM) for producing selected-response test items. Building on our initial study comparing prompt-engineered LLM-generated items to human-generated items, a second study was conducted to measure the impact of fine-tuning LLM using operational test data to generate items across two batches.

3. **Uncovering Ability Trajectories in Longitudinal Assessment: A Growth Mixture Modeling Approach**
   Hahyeong Kim, University of Illinois at Urbana-Champaign; Nikole Gregg, National Commission on Certification of Physician Assistants; Yanlin Jiang, NCCPA; Christiana Akande, National Commission on Certification of Physician Assistants

   Using two years (eight-quarter) longitudinal recertification assessment data, we explored ability trajectories with Growth Mixture Model and found three-class quadratic solution fits best. Most examinees showed “Recovery” (early decline, then rebound); smaller groups were “High Stable” and “Improvers.” Hours worked predicted “High Stable”; age and lower confidence predicted “Improvers.”

4. **Using Simulation-Extrapolation Bayesian Logistic Regression to Calibrate Pretest Items in Multistage Testing**
   TsungHan Ho, ETS

   Bayesian logistic regression (BLR) is a method developed to address the response sparseness of linking items, which undermines the quality of the link to the item bank in multistage testing. To minimize the Bayesian bias-variance trade-off, a Simulation-Extrapolation BLR method is proposed to improve its performance, evaluated using simulation data.

5. **IRT Pre-smoothed Frequency Estimation (FE) Equating Method**
   Min Liang, National Board of Osteopathic Medical Examiners (NBOME); Won-Chan Lee, University of Iowa; Huan Liu, Riverside Insights; Haimiao Yuan, Prometric LLC; Jinghua Liu, The National Board of Osteopathic Medical Examiners; Stuart Barnum, National Board of Osteopathic Medical Examiners

   This study aims to propose a new IRT-based Frequency Estimation (FE) equating method under the CINEG design and compares its performance with seven widely used equating methods through a simulation study. Results indicate that the IRT FE method can reduce the standard error of equating.

6. **Equating outcome comparison using different stability detection methods in vertical scale development**
Hongwook Suh, Cambium Assessment

This study compares multiple item parameter drift detection methods with restricted applications for linking items in developing a vertical scale. Simulated data from a statewide Grade 3-8 mathematics test are used to evaluate their impact on expected equating outcomes, such as grade-to-grade growth, variability, and ALD group proportions.

7. **Advancing Language Assessment Through Vertical Scaling: Psychometric Insights from WIDA's Assessment**
Kyoungwon Bishop; Hacer Karamese, WIDA at the University of Wisconsin-Madison; Sooyong Lee, WIDA at the University of Wisconsin-Madison

This presentation shares findings from WIDA's vertical scaling studies in Speaking and Writing across Grades K–12. It outlines score scale development for longitudinal tracking, calibration strategies, psychometric challenges, and empirical results. Implications for operational use and future research are discussed to support English language development measurement.

8. **Optimal Vertical Scaling Design for Short Tests with Polytomous Items**
Sooyong Lee, WIDA at the University of Wisconsin-Madison; Kyoungwon Bishop, University of Wisconsin - Madison; Hacer Karamese, WIDA at the University of Wisconsin-Madison

This study examines vertical scaling for very short polytomous tests. Simulations under the Rasch Rating Scale Model show that at least three common items, or half of available anchors, are needed to achieve stable linking. Anchor coverage, not step-value range, drives bias control, offering guidance for writing assessments.

9. **Polytomous Item Drift in Pretesting: Challenges for Next-Generation Test Transition**
Lucia Liu, Ascend Learning; Kari Hodge, Ascend Learning; Xuechun Zhou, Ascend Learning

This study investigates parameter drift in polytomous items pretested within dichotomous tests. Using IRT and WRMSD comparisons, we found consistent shifts in item parameters post-release. These shifts impact equating and score validity, highlighting the need for modified pretesting practices in mixed-format assessment programs.

10. **Impact of Item Drift on Pre-Equated Scores: A Rasch-Based Investigation**
Eric Asare, Old Dominion University; Daniel Edi, NC Department of Public Instruction

This simulation study examines how item drift affects score accuracy in pre-equated assessments. Monte Carlo simulations revealed that item drift can misclassify 42–61% of students, even with fixed item parameter calibration. Findings underline the need for robust detection methods to mitigate item drift in high-stakes testing.

11. **Detecting Drifted Items in Scale Linking with Machine Learning Techniques**
Tong Wu; Hyeonjoo Oh, Riverside Insights; Stella Kim, University of North Carolina Charlotte; Hsin-Ro Wei, Riverside Insights

This study applies machine learning-based anomaly detection approaches to identify outliers in common items in scale linking. It compares these approaches with traditional methods under diverse simulation conditions, which provides insights into their relative performance and effectiveness in enhancing the accuracy and robustness of linking procedures.

12. Evaluating Sampling Strategies for Random Equivalent Groups in Large-Scale Assessments
Guangyun Liu, NBOME; Ying Lu, The College Board; Amy Hendrickson, The College Board

This study investigates alternative sampling strategies for Random Equivalent Group equating in large-scale assessments. Using simulation across multiple subjects, we compare random cluster sampling, stratification, and performance-based approaches. Findings suggest random cluster sampling provides the most representative and equivalent samples, raising important considerations for operational equating practices in educational measurement.

13. Equipercentile Equating Precision Across Sample Size and Test Length
Talal Alzabidi, University of North Carolina Greensboro

Equipercentile equating accuracy depends on sample size and test length. A full-factorial Monte Carlo (N=500–3000; 20 vs 40 items) shows negligible bias but variance-dominated error: increasing N lowers RMSE, longer tests raise it at fixed N. We derive planning curves specifying minimum N by length for target accuracy.

14. Visualizing Knowledge Gaps Using Semantic Latent Space Models
William Muntean, National Council of State Boards of Nursing; Joe Betts, National Council of State Boards of Nursing

This study introduces a latent space item response theory model that incorporates semantic embeddings of exam items to produce dual diagnostic scores: traditional performance and estimated conceptual understanding. Visual radar reports reveal knowledge gaps, guiding targeted remediation and transforming summative high stakes test data into formative feedback for exam takers.

15. A Framework for Quantifying the Effect of Item Compromise on Linking Functions
Aaron Myers, American Board of Internal Medicine; Neil Dorans, Independent Consultant

Quantifying the effect of item compromise on IRT-based linking is crucial for ensuring test fairness. An analytical framework is proposed to estimate the bias in linking coefficients as a function of compromised items and examinees with preknowledge. This framework helps test practitioners make evidence-based decisions about test security breaches.

16. A Systematic Review of Machine Learning-Based Approaches for Detecting Test Fraud
Daihui Xiao; Kylie Gorney, Michigan State University

This systematic review synthesizes several studies that apply machine learning techniques—including supervised, semi-supervised, and unsupervised learning—to detect test fraud. By identifying strengths, limitations, and research gaps, we hope to provide an overview of the methodological landscape and help inform the development of data-driven approaches to ensuring test security.

17. Methodological Practices in Educational and Psychological Test Translation: A Systematic Review
Jack Riley, University of Nebraska - Lincoln; Jordan Wheeler, University of Nebraska - Lincoln

This systematic review examines methodological practices in test translation across 93 quantitative and 18 qualitative studies. Findings reveal inconsistent reporting of validation methods, inadequate justification of procedures, and variation in practices within identical measures. Results highlight gaps in the standardization of protocols, providing recommendations for improving test translation methodology.

18. SAT Scaled Text Complexity with Item Response Theory and Natural Language Processing
Sunhee Kim; Judit Antal, College Board

Psychometric and language models are integrated to quantify passage complexity on the SAT scale. Using IRT-based probability criterion mapped to the SAT scale, we trained ML-models to predict passage difficulty from textual features and further derive interpretable sub-scores via factor analysis. Results show strong alignment with observed operational statistics.

19. **Comparability of Virtual and In-Person Proctoring for State Assessments**
Timothy O'Neil, Pearson

This study evaluated the comparability of virtually proctored and in-person administered state mathematics assessments for online students. Results indicated no significant differences in performance or item functioning, supporting remote proctoring under ESSA for large-scale state summative assessments, with implications for policy and practice.

20. **Does the Shortened GRE General Test Score Predict Graduate School GPA?**
Guangming Ling, ETS; Yuan Wang, ETS

We examined the predictive validity of the shortened GRE General Test by analyzing its correlation with first-semester graduate GPA, using data from 920 students across 253 institutions. Low but positive associations overall were found, with variations by gender, major area, and citizenship, consistent with earlier findings for the original GRE.

Innovative Partnerships for Indigenous Language Sustainability and Culturally Grounded Assessments

Coordinated Paper Session
1:45 PM – 2:45 PM
Intercontinental Los Angeles Downtown, Floor 5 : Silver Lake B

Western assessment practices have long marginalized Indigenous students by privileging dominant cultural norms and English-only frameworks, often undermining sovereignty and erasing local knowledge systems. This symposium highlights a new generation of assessment research that is Indigenous-led, co-designed, and rooted in sustaining language and culture. Together, the three papers demonstrate that building fair, valid, and culturally meaningful assessments requires re-centering authority with Indigenous nations and engaging in long-term, trust-based partnerships.

The first paper shares findings from an NCME-funded project that reconceptualizes validity as co-design work, positioning fairness and cultural integrity as governing principles rather than procedural add-ons. The second paper illustrates how Indigenous-led partnerships, grounded in respect, reciprocity, and responsibility, can produce culturally sustaining assessments that both honor sovereignty and generate actionable insights for educators and policymakers. The third paper details the collaborative development of curriculum-based reading measures in Alaska Native languages, showing how local educators, linguists, and Elders shaped every stage of the process from construct definition to field testing.

Together, these projects demonstrate pathways for transforming assessment from a tool of assimilation into a practice of affirmation, sustaining Indigenous languages, strengthening cultural continuity, and modeling approaches that are transferable to other currently and historically marginalized communities.

Presentations:

1. **Co-Design as Validity Work: Advancing Fair, Community-Owned Measurement for Language and Culture**
Pohai Kukea Shultz, University of Hawaii - Manoa; Staci Block, California Indian Education for All; Edynn Sato, Sato Education Consulting LLC; Vitaliy Shyyan, WIDA, University of Wisconsin - Madison

2. **Collaborative Pathways for Indigenous-Led Assessment Research: Building Sustainable Partnerships for Language and Cultural Revital**
Staci Block, California Indian Education for All; Vitaliy Shyyan, WIDA, University of Wisconsin - Madison; Kyungjin Sohn, WIDA, University of Wisconsin - Madison

3. **Co-Design of Curriculum-based Measures of Reading in Alaska Native Languages**
Gina Biancarosa, University of Oregon; Katherine Wright, University of Oregon; Sylvia Linan-Thompson, University of Oregon

Validating Literacy Screening Profiles to Identify Risk for Reading Difficulties and Dyslexia

Coordinated Paper Session
1:45 PM – 2:45 PM
Intercontinental Los Angeles Downtown, Floor 5 : Boyle Heights

The authors examine the validity of early literacy profile classifications that were developed to screen students in kindergarten through grade 3 for reading difficulties and characteristics of dyslexia. Operationally, screening results are intended to be used for the purpose of identifying students about whom more information is needed to determine their next steps for intervention or diagnostic testing. The first paper discusses the processes and outcomes of the standard setting and profile validation studies that were used to establish the screening profiles. The second paper examines the variance explained in reading outcomes by screening classifications and profile components. The third paper applies latent profile analysis to identify distinct profiles of reading difficulties for kindergarten through grade 3 students. The results of these studies support a deeper understanding of the utility and defensibility of using this profile approach to identify student risk classifications in a school setting.

Chair:
Michelle Boyer

Discussant:
Jessalyn Smith (DRC)

Presentations:

1. **Establishing and Validating Performance Standards for Universal Early Literacy**
   Ricardo Mercado, DRC; Michelle Boyer, Data Recognition Corporation; Sara Kendallen, DRC; Kara Courtney, DRC; Julie Pointer, DRC

2. **Using Screening Profile Classifications and Components to Explain Reading Outcomes**
   Aria Immanuel, University of Massachusetts Amherst; Michelle Boyer, Data Recognition Corporation; Kara Courtney, DRC

3. **Identification of Subgroups of Early Readers**
   Christine DiStefano, University of South Carolina; Huijuan Wang

Applications of Network Analysis

Individual Paper Session
1:45 PM – 3:15 PM
Intercontinental Los Angeles Downtown, Floor 5 : Westwood

Chair:
Yuan-Ling Liaw (IEA Hamburg)

Discussant:
Evelyn Johnson (Riverside Insights)

Presentations:

1. **Community Detection in Nominal Ising Networks: Unveiling Latent Patterns**
   Ae Kyong Jung, University of Iowa; Jonathan Templin, University of Iowa

   This proposal introduces a novel approach to uncovering student misconceptions from multiple-choice responses using psychometric network analysis. By examining the relationships among indicators and applying a network community detection algorithm, the method identifies clusters of misunderstandings, offering deeper insights into students' learning patterns and enhancing diagnostic assessment.

2. **Comparing Parallel Analysis and Exploratory Graph Analysis in Large Factor Model with Binary and Ordinal Data**
   Ruoqian Wu, University of Illinois Urbana-Champaign; Xinchang Zhou; Yan Xia, University of Illinois at Urbana-Champaign

   This simulation compared Parallel analysis was conducted using different correlation matrices (Pearson, polychoric) and extraction methods (PCA, PAF, MRFA), and exploratory graph analysis (Glasso, TMFG) with various community detection algorithms(Louvain, fast greedy, and walktrap) in large factor models using binary and four-category data.

3. **Examining Teacher Beliefs and Perceptions for Creativity using Network Analysis**
   Shahnaz Safitri; Nielsen Pereira, Purdue University

   This study investigated the interrelations between teachers' beliefs and perceptions of creativity via network analysis. Growth mindset and creative self-efficacy were identified as central drivers of creative teaching. The findings demonstrate the utility of network analysis for modeling complex educational constructs and guiding targeted interventions in teacher professional development.

4. **Network-Based DIF Detection of Multiple Grouping Variables in Polytomous Data**
   Kuan-Yu Jin, Hong Kong Examinations and Assessment Authority; Chi-Chen Chen, National Academy for Educational Research

   The exploratory graph analysis (EGA) was introduced as an intuitive and effective method for detecting differential item functioning (DIF) in polytomous assessments with multiple grouping variables. Simulation results showed that EGA can accurately and quickly identify DIF items, showing its comparability and advantages over the multiple-indicator-multiple-cause (MIMIC) method.

5. **Bayesian Graphical Models in Factor Analysis: A Prior Comparison**
   Yifan Zhang, The University of Hong Kong; Jinsong Chen, The University of Hong Kong

   Graphical lasso has been applied in partially confirmatory factor analysis (PCFA; Chen, 2020) to model residual covariance. We extend PCFA by embedding Bayesian graphical modeling to estimate sparse factor correlations. Three priors (lasso, horseshoe, spike-and-slab) are used to compare regularization of residual and factor structures through simulation and empirical studies.

Fairness Concerns with AI

Individual Paper Session
1:45 PM – 3:15 PM
Intercontinental Los Angeles Downtown, Floor 7: Roosevelt B

Chair:
Daniel Oyeniran (The University of Alabama)

Discussant:
Kyla McClure (University of Colorado Boulder)

Presentations:

1. **Ethical Issues and Artificial Intelligence (AI) within an Educational Context**
   Lisa Clark, City University of New York

   The focus of this presentation is how to harness benefits of artificial intelligence while maintaining humanity? It includes an analysis of six artificial intelligence projects in undergraduate teacher preparation programs. The foundation is an examination of the application of Artificial Intelligence through the lens of the philosophy of ethical behavior.

2. **Utility of Off-the-Shelf Generative AIs for DIF Detections**
   Chalie Patarapichayatham, HMH; Gozde Sirganci, Southern Methodist University; Akihito Kamata

   This study explores the use of publicly available generative AIs for DIF detection without additional training and/or fine-tuning. We also examine the agreement on DIF detection and content bias reviews (CBR) between the AIs and subject matter experts (SME).

3. **Exploring Justice-Oriented Frameworks for Evaluating LLM Parameter-Induced Bias Patterns**
   Lauren White, Pearson; Michael Chajewski, Pearson; Sarah Quesen, WestEd

   This exploratory study examines how parameter configurations in large language models may influence bias in educational contexts. Using a justice-oriented theoretical framework, we analyze a focused sample of LLM responses across select temperature and top-p settings, providing preliminary evidence and guidance for practitioners seeking to minimize bias in AI applications.

4. **When Testing Takes Longer: Item Features and Latency in Statewide Science Assessment**
   Yi-Fang Wu, Cambium Assessment, Inc.

   By integrating psychometric analysis, NLP-derived features, and ML/LLM methods, this study investigates how item characteristics affect response time in a statewide science assessment and whether increased testing time in the post-pandemic cohort reflects literacy demands. Subgroup analyses explore whether these patterns differ across student groups, highlighting potential equity implications.

5. **AP Teachers' ChatGPT Use: A Cluster-Based Mixed-Methods Investigation**
   Dan Song, The University of Iowa; Frederick Poole, Michigan State University

   This study builds on our prior research using the TPACK framework to examine AP teachers' use of ChatGPT. We conducted cluster analysis to identify three teacher groups and applied qualitative analysis to the group reporting the lowest use to explore potential reasons for resisting ChatGPT integration in AP courses.

Frameworks and Approaches for DCM

Individual Paper Session
1:45 PM – 3:15 PM
Intercontinental Los Angeles Downtown, Floor 5 : K-Town

Chair:
Nicolas Mireles

Discussant:
Madeline Schellman (Pearson)

Presentations:

1. **Teamwork Cognitive Diagnostic Modeling**
   Peida Zhan, Zhejiang Normal University; Zhimou Wang, Zhejiang Normal University; Beijing Normal University; Gaohong Chu, Zhejiang Normal University; Haixin Qiao, Zhejiang Normal University

   This study proposes a teamwork cognitive diagnostic modeling framework comprising 12 specific models which are designed to capture the interdependence among team members through emergent team cognitions by jointly modeling individual cognitive attributes and a team-level construct, termed teamwork quality, which reflects the social dimension of collaboration.

2. **Nonparametric Diagnostic Classification in Ordinal Response Items**
   Joemari Olea, University of Texas at Austin; Hyeon-Ah Kang, University of Texas Austin

   The study introduces a nonparametric approach to cognitive diagnosis that accommodates general ordinal responses. We present a saturated modeling framework capable of handling both dichotomous and polytomous item formats. A simulation study and analysis of real data are conducted to demonstrate the capability of the proposed method.

3. **Extra-Dispersion and Robust Errors in Cognitive Diagnostic Models**
   Michel Cordoba Perozo, Purdue University

   This paper introduces a Bayesian semiparametric framework to obtain robust standard errors for binary cognitive diagnostic models. By addressing extra-dispersion overlooked in traditional estimators, the approach improves parameter inference and supports more reliable latent classification. Simulation studies demonstrate technical gains, while an application illustrates practical relevance.

4. **Extending Cognitive Diagnosis: A Latent Space Item Response Model Approach**
   Brian Harrold, University of South Carolina; Brian Habing, University of South Carolina

   A synthesis of the Loglinear Cognitive Diagnosis Model (LCDM) and the Latent Space Item Response Model (LSIRM) is introduced. The unified framework combines diagnostic classification with latent space modeling of residual dependence, enabling detection of item and respondent relationships that aren't computed by the specified cognitive diagnostic models.

5. **Accelerating Mastery Promotion via Information Gain-guided Item Selection**
   Sangbeak Ye, Florida Atlantic University

   We propose an adaptive mastery-detection framework for cognitive diagnosis (e.g., G-DINA) that jointly optimizes item selection and stopping. Detection is implemented via either CUSUM or posterior-probability thresholds; a non-adaptive WCIS baseline provides comparison. Simulations with heterogeneous items and transition variability demonstrate accurate control of false alarms and reduced delays.

Methodological Issues in Large-Scale Assessment

Individual Paper Session
1:45 PM – 3:15 PM
Intercontinental Los Angeles Downtown, Floor 7: Roosevelt A

Chair:
Mohammed Abulela (MetaMetrics, Inc. and University of Minnesota)

Discussant:
Mohammed Abulela (MetaMetrics, Inc. and University of Minnesota)

Presentations:

1. **Comparison of the IRT-C and E-CDM to Detect and Explain DIF**
   Kevin Krost

   Differential item functioning (DIF) was evaluated between English- and Spanish-speaking students on released science items from the 2011 TIMSS using the IRT-C and E-CDM models. Several items exhibited DIF, and covariates were able to explain the DIF. Last, item content features explained the remaining DIF after modeling the covariates.

2. **Measurement Invariance of PISA Curiosity Scale Across Public and Private South-American Schools**
   Solange Barros-Bustos, The University of Kansas; Sean Joo, University of Kansas

   This study tested measurement invariance of PISA 2022 Curiosity scale across public-private schools in seven South American countries. Two poorly performing items were removed, resulting in a revised 8-item model with acceptable fit. Multiple-Group Confirmatory Factor Analysis supported strict invariance, with changes in fit indices within recommended thresholds (Chen, 2007).

3. **Variations in DIF Detection Due to SES Representation: Causal Inidcators vs Composites**
   Alejandro Martinez, University of North Carolina at Chapel Hill; Nicolas Mireles

   We examine how modeling Socioeconomic Status influences Differential Item Functioning conclusions. Using PISA 2022 science data and a follow-up simulation, we contrast composite, causal, and effect indicator specifications under various conditions. Results show detection and effect sizes depend on modeling, offering guidance for stronger interpretations and decisions in DIF analyses.

4. **Longitudinal Inference without Longitudinal Data: A Copula-Based Unification of TAMP and SGP**
   Damian Betebenner, Center for Assessment; Henry Braun, Boston College

   Many large-scale assessments (e.g., NAEP or TIMSS) are cross-sectional, precluding traditional student-level growth analysis. We propose a copula-based approach to modeling of dependence structure between separate score distributions to simulate pseudo-longitudinal cohorts. The approach yields policy-relevant growth estimands (e.g., state/country growth rankings) without tracking individuals. The approach holds promise to transform assessment reporting.

5. **Assessing Practical Measurement Equivalence in Nested Data: A Bayesian ROME Extension to Multilevel CFA**
   Yichi Zhang, Georgia Institute of Technology

   This study extends the Bayesian Region of Measurement Equivalence (ROME) framework to nested data, introducing an approach for quantifying measurement noninvariance across clusters within two-level confirmatory factor analysis models. Using the 2007 TIMSS data, we illustrate how the method identifies school-level differences and supports fairness evaluations in large-scale educational assessments.

Psychometric Overlord Auditions Take 2: Prove Me Wrong

Coordinated Poster Session
1:45 PM – 3:15 PM
Intercontinental Los Angeles Downtown, Floor 7: Hollywood Ballroom II

One of the most critical aspects of any successful NCME conference is the exchange and debate of ideas among NCME members. The 2026 Psychometric Overlord Auditions poster session will be your opportunity to discuss, debate, and evaluate controversial ideas from leaders across the measurement community. The winners from the 2025 Psychometric Overlord competition return and new challengers have emerged ready to share their best ideas for how to shake-up the measurement community.

All speakers will be asked to share an idea or claim they have that contradicts current thinking in our field and challenge the session attendees to “prove me wrong”.

Join us for this session and get ready to discuss the innovative, intriguing, and possibly off-the-wall claims of your colleagues. 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:
Susan Davis-Becker (ACS Ventures, LLC) and Andrew Wiley (ACS Ventures, LLC)

Presentations:

1. **Norm-referenced testing that just rank orders students is one of the main reasons people hate tests**
   Kristen Huff, Curriculum Associates

2. **AI will never replace field testing with actual students**
   Susan Lottridge, Pearson

3. **Fairness is a design choice, not a psychometric outcome**
   Ye Tong, National Board of Medical Examiners

4. **There is no such thing as consequential validity**
   Andrew Jones, American Board of Surgery

5. **For as long as I shall live, I will be convinced that nobody actually understands what a logit is — including the people building AI models with them; prove me wrong**
   Lisa Keller, University of Massachusetts - Amherst

6. **You can reduce Contamination By Tolerating Contamination**
   James Wollack, University of Wisconsin - Madison

7. **We should not stop test-takers from cheating**
   Amy Hendrickson, The College Board

8. **Section Time Limits Don’t Control Item-Level Speededness**
   Kyung (Chris) Han, GMAC

9. **State K-12 testing programs will continue to test every child every year and never adopt any kind of sampling approach**
   Bradley McMillen, Wake County Public School Sustem

10. **The public’s negative experiences of assessments in schools leads to the public’s mistrust of certification, licensure, and pre-employment assessments.**
    Amin Saiar, Board of Pharmacy Specialties

Beverage Break
Social
2:45 PM – 3:45 PM
Intercontinental Los Angeles Downtown, Floor 5 : Wilshire Grand Prefunction

Hot coffee & tea, lemonade, and iced tea will be available in the Wilshire Grand Prefunction area.

Automated Scoring: Combining and Comparing Human and AI Raters
Individual Paper Session
3:30 PM – 5:00 PM
Intercontinental Los Angeles Downtown, Floor 7: Roosevelt A

Chair:
Mike Maksimchuk (Kent Intermediate School District)

Discussant:
Edward Wolfe (Iowa Testing Programs / University of Iowa)

Presentations:

1. **Hybrid Human-AI Rubric Development for Interpretable Automated Evaluation in Computer Science Assessments**
   Warren Li, Associate Research Scientist; Daisy Wise Rutstein, edCount, LLC; John Whitmer, Learning Data Insights

   This presentation describes the application of LLMs to evaluate student coding tasks. We discuss the initial development of rubrics and processes to enable accurate automated scoring. Analysis resulted in high human-AI agreement for some features. The process revealed the potential and limits of LLMs for pedagogical utility.

2. **Integrating Human and LLM-Based Scoring: A Cross-Classified IRT Approach**
   Eun Hye Ham, Kongju National University; Yun-Kyung Kim, UCLA

   This study applies cross-classified IRT models to examine the comparability of human and LLM-based scorings of constructed responses. By modeling rater variability across prompt types and scoring criteria within a unified psychometric framework, the approach provides fine-grained analyses of severity and consistency, clarifying the strengths and weaknesses of LLM scoring.

From Validity Arguments to Validity Synthesis and Judgment

Coordinated Paper Session
3:30 PM – 5:00 PM
Intercontinental Los Angeles Downtown, Floor 5 : K-Town

According to the Standards, validity represents the degree to which evidence and theory support the interpretations and uses of test scores. Yet in practice, operationalizing argument-based approaches to validity has proven unwieldy: frameworks often become bloated, jargon-heavy, and disconnected from decision-making that could lead to actionable improvements to testing programs. This coordinated session presents a practical framework for moving from validity arguments to a validity judgment. Drawing on work with the Multi-State Alternate Assessment (MSAA), the session will make conceptual and applied contributions. Bethany Spangenberg will provide some history and context for the alternate assessments produced by MSAA. Derek Briggs will motivate the need for a new approach to validity arguments. Mike Russell will present a key piece of framework, which organizes the central assumptions underlying test score interpretations into coherent categories and develops a crosswalk between this evidence and chapters in a technical report. Frank Palladero will reflect on efforts to use this organizational structure to synthesize evidence with each central assumption. Chris Domaleski will share his perspective as an independent evaluator applying the validity judgment framework in practice. Scott Marion will serve as discussant, situating the work within the broader context of validity theory and practice.

Chair:
Derek Briggs

Discussant:
Scott Marion (Center for Assessment)

Presentations:

1. **Background on the MSAA Consortium Context**
   Bethany Spangenberg, Deputy Associate Superintendent of Assessment

2. **Making Validity Arguments Practical**
   Derek Briggs, University of Colorado - Boulder

3. **Organizing Validity Evidence**
   Michael Russell, Boston College

4. **Synthesizing Validity Evidence into Arguments**
   Frank Padellaro, Cognia

5. **Third Party Evaluations and Judgments of the Validity Synthesis**
   Christopher Domaleski, Center for Assessment

Inclusion, Equity, and Fairness in TIMSS and PIRLS

Coordinated Paper Session
3:30 PM – 5:00 PM
Intercontinental Los Angeles Downtown, Floor 5 : Westwood

Large-scale assessments aim to provide internationally-comparable results that generalize to diverse student populations worldwide. Growing diversity in student populations and languages across participating countries, combined with the transition to digital platforms, renews emphasis on issues of exclusion, accessibility, and bias that can undermine the validity of results. While digital platforms enable integrating modern assessment technologies, they require addressing new threats to inequity to maintain representativeness and credibility. This session features four studies conducted to address inclusion, equity, and fairness in the design, administration, and analysis of TIMSS and PIRLS. Research examines the impact of excluding students with disabilities or language barriers from samples and discusses strategies to promote inclusiveness. Issues related to digital accessibility are investigated, focusing on evaluating typing demands across different writing systems. Moreover, a framework is provided for detecting and addressing language bias in scoring with artificial intelligence (AI), illustrating approaches to evaluate fairness in AI-based processes. Lastly, based on research examining the impact of nonresponse on measures of socioeconomic status, the session informs approaches to examine inequities within and across education systems. Together, the research highlights the trade-offs between methodological rigor and fairness, and provides evidence-based strategies to enhance the validity of international assessment results.

Chair:
Bethany Fishbein (Boston College)

Discussant:
Irini Moustaki (London School of Economics and Political Science)

Presentations:

1. **Understanding and Addressing Student Exclusions in TIMSS and PIRLS**
   Umut Atasever, IEA; Sabine Meinck, IEA; Bethany Fishbein, Boston College; Matthias von Davier, Boston College

2. **Exploring the Impact of Typing Demand on Assessment Performance Across Languages**
   Erin Wry, Boston College; Bethany Fishbein, Boston College; Matthias von Davier, Boston College

3. **Addressing Fairness in AI Scoring of Written Responses Across Languages**
   Ummugul Bezirhan, Boston College; Ji Yoon Jung, Boston College; Matthias von Davier, Boston College

4. **Capturing Home Resources and Opportunity to Learn in International Assessments: Insights from PIRLS 2021**
   Dihao Leng, East China Normal University; Deepthi Kodamala, Boston College; Katherine Reynolds, Boston College

Individual eBoards: Classroom & Formative Assessment, Score Reporting, Behavioral Patterns

Electronic Board Session
3:30 PM – 5:00 PM
Intercontinental Los Angeles Downtown, Floor 5 : Hancock Park

Presentations:

1. **Investigating Exam Timing Effects in High-Stakes Testing**
   Xiaoting Zhong, University of Iowa; Hanwook Yoo, Ascend Learning

   This study examines fairness in high-stakes testing by investigating time-of-day effects using a regression discontinuity design. Results reveal a significant performance discontinuity between two groups (morning vs. afternoon), with examinees tested afternoon scoring slightly higher. Findings demonstrate rigorous causal methods to advance educational measurement practice related to potential test fairness.

2. **Principals as Assessment Leaders Across Rural and Urban Contexts**
   Cristyne Hebert, Associate Professor; Jaclyn Roach, University of Regina

   This study examines how principals in rural, urban, and remote schools enact assessment leadership. Interviews with 22 principals were analyzed using Charteris and Smardon's (2022) framework, revealing five shared leadership dimensions enacted differently by context. Findings demonstrate the importance of context-sensitive support for strengthening teacher assessment literacy and practice.

3. **Planning for the Use of the Formative Assessment Process**
   Tara Kintz, Michigan Assessment Consortium; John Lane, Michigan Assessment Consortium; Edward Roeber, Michigan Assessment Consortium

   This session investigates the central role of planning in the formative assessment process. Findings illustrate how intentional planning strengthens educator practice, fosters student agency, and advances socioculturally responsive assessment. Implications for professional learning, teacher preparation, and policy are discussed.

4. **Understanding Purpose-setting Validity in Upskilling Formative assessment for IA via AI**
   Li Liang

   Given the significance that validity of purpose-setting is the priority factor in upskilling formative assessment for Human intelligence augmentation (IA) via artificial intelligence (AI), this article aims for an understanding how purpose-setting validity can better inform teaching and upskilling formative assessment for human IA via AI.

5. **Scaffolding Argumentative Writing with Multiple Perspectives in Middle School**
   Yi Song, ETS Research Institute; Chunyi Ruan, ETS; Michael Suhan, Educational Testing Service

   This study explored an intervention designed to improve essay planning skills of middle school students. We found the intervention improved student awareness of the alternative perspective in their plans. Students who used the critical discussion or persuasion templates outperformed their peers in both planning and final essays.

6. **Detecting Intuitive and Deliberative Thinking Modes in PISA 2022 Creative Thinking Assessment**
   Xiaoxiao Liu, University of Alberta; Yaxin Dong, university of alberta; Bin Tan, University of Alberta; Ying Cui, University of Alberta; Okan Bulut, University of Alberta

   Applying a Hidden Markov model to analyze the time to first action on creative thinking tasks, this study identified two distinct thinking modes: intuitive and deliberative. High-performing students spent more time and deliberately chose their mode, while low-performing students spent less time and switched between modes more randomly.

7. **Extreme Timing Behaviors and Ability Estimation in the PIAAC Using IRTree Approach**
Yulim Kang, Yonsei University; Sohee Kim, University of South Alabama; Seulee Lee

This study analyzes PIAAC numeracy assessment logs to detect extreme timing behaviors such as rapid guessing and over-deliberation using Item response Tree (IRTree) models. By integrating response times with accuracy data, it demonstrates how process data enhances ability estimation compared to traditional IRT models relying solely on correct/incorrect responses.

8. **A Hidden Markov Approach to Understand Behavior Transitions in Test-Taking Processes**
Ruiting Shen, New York University; Klint Kanopka, New York University

Few previous studies of process data focus on the evolution of respondents' behavior across the entire test. Using Hidden Markov Models applied to NAEP data, we identify four hidden behavioral states. By tracking transitions between states across items, we describe response process trajectories over the course of the test.

9. **Process Data in Educational Assessment: A Systematic Review**
Ya Mo; Artemis Rainn, Boise State University; Tasnubha Bably, Boise State University

This paper presents a systematic review of 85 studies that utilize process data in educational assessments. Studies are classified by data type, subject area, analytical methods, and outcomes. Findings reveal key trends in response-time and action-based measures, with implications for assessment design, validity, and future applications of process data analytics.

10. **How Well Does the 2PL Model Fit Non-Cognitive Testing Data?**
Wenhao Wang, HumRRO; Corissa Rohloff, HumRRO; Allen Goebl, HumRRO

Non-cognitive data are collected to assess social and emotional learning outcomes. We aimed to study how well the two-parameter logistic (2PL) model fits non-cognitive data by fitting the model to ideal point model data. The results indicate that the 2PL model has poor item-level model fit to non-cognitive data.

11. **Leveraging Assessment Metadata to Capture Foundational Executive Functioning Skills**
Sophie Litschwartz, MDRC; Emily Hanno, MDRC; Victor Porcelli, MDRC; Emily Swinth, MDRC

Traditional executive function (EF) assessments for preschoolers are expensive and often lack ecological validity. We use response time metadata from preschoolers' tablet based assessments (N=268) to develop new EF measures. Our measures achieved reliabilities between .5-.7 and correlated .3-.7 with standardized EF assessments, supporting cost-effective, ecologically valid early childhood EF measurement.

12. **Beyond Testing: Incorporating Culturally Responsive Assessments in Informal Education**
Jayma Koval, Georgia Institute of Technology; Diley Hernández, Georgia Institute of Technology

This study examines the use of culturally responsive assessments within an informal computer science program for Latinx students. Using qualitative analysis, findings illustrate how formative and performance-based summative assessments aligned with culturally responsive practices supported student learning. Results highlight challenges, perceptions, and implications for designing assessment in informal settings.

13. **Improving Standard Setting Resources to Promote Better-Informed Score Uses**
Francis O'Donnell, National Board of Medical Examiners; Leslie Keng, National Board of Medical Examiners

This study demonstrates a targeted approach to improve standard setting reports in the context of a program that gives users a choice of which cut scores to use. The approach is informed by user-centered needs assessment, usability, and alignment to professional standards, with the goal of supporting better-informed score uses.

14. Early Identification of At-Risk Students Using ROC Analysis of Interim Assessments
Yuming Liu, Cambium Assessment Inc

This study uses ROC analysis to evaluate how well fall and winter interim assessments predict spring summative proficiency in Grades 4–8 ELA and Math. Results show strong classification accuracy (AUC = 0.89–0.92), supporting the use of interim scores for early identification and instructional intervention. (46 words)

15. Leveraging longitudinal data for comparability of interim assessments scores across administration conditions
Luciana Cancado, Curriculum Associates; Montserrat Valdivia Medinaceli, Curriculum Associates; Logan Rome, Curriculum Associates

This study explores the use of longitudinal data to provide evidence of score comparability across administration conditions of a K–12 interim assessment. Models incorporating prior student scores are used to predict future scores, and aggregated deviations from observed scores used to evaluate comparability without needing a control group.

16. A Randomized Control Study of Extended Time Accommodations
Ramsey Cardwell, Duolingo; Jill Burstein, Duolingo; William Belzak, Duolingo; Ping-Lin Chuang, Duolingo

Randomizing 8,988 DET practice test takers to standard or +50% time, we estimate extended time effects overall, by task type, and by self-reported condition. Gains exceed error for some groups (e.g., ADHD, hearing) and writing tasks broadly. Findings partly support the maximum potential thesis and suggest utility of tailored accommodations.

17. Embedded Text-to-Speech in Testing: Usage Patterns and Performance Effects
Yoonjeong Kang, Cambium Assessment; Sarina Bridges, Cambium Assessment

This study examines how students use embedded text-to-speech (TTS) in large-scale computer-based assessments and its impact on performance. Using propensity score matching, we compare eligible TTS users and non-users in ELA and mathematics. Findings highlight variable usage patterns and performance benefits for students with disabilities.

18. Validity Evidence for Responses to Feedback: A Theory of Action
Piet Wesling, University at Albany, SUNY; Angela Lui, City University of New York; Joseph Garofalo, University at Albany, SUNY; Diana Akhmedjanova, National Research University Higher School of Economics, Moscow, Russia; Heidi Andrade, University of Albany

This study examines validity evidence for formative feedback in a diagnostic assessment of self-regulated learning. Using a theory of action framework, we analyzed interpretive claims and student responses captured through think-aloud protocols and interviews. Findings highlight supported claims, unintended consequences, and implications for designing feedback systems in educational assessments.

19. Post-Pandemic Shifts in College Readiness: What is the Impact on Validity Coefficients over Time?
Jessica Marini, College Board; Emily Shaw, College Board

COVID-era performance shifts have raised questions about how well high school academic measures can continue to predict college performance. This study examines high school grade point average (HSGPA) and test score validity coefficients with college GPA, pre- and post-pandemic. Findings show that these predictive relationships remain stable despite performance disruptions.

20. **Validity Evidence for Qualitative Data Analysis in Educational Research: Human vs. Machine Coding**
Gamze Karaer, Washington State University; Shenghai Dai, Washington State University; Yuliya Ardasheva, Washington State University; Anne Guerrettaz, Washington State University; Yun-Ju Hsiao, Washington State University; Lindsay Lightner, Washington State University; Danica Garcia, Washington State University; Adam Coldiron, Washington State University

This study evaluates validity, reliability, and scalability of qualitative coding by comparing human and machine approaches. Using interviews, focus groups, and observations, we identified systematic differences across cases and fair-to-moderate alignment (κ = 0.36) between machine-generated and human-coded themes. Results highlight the promise and limitations of computational methods for qualitative-research.

Properties of Specific Tests

Individual Paper Session
3:30 PM – 5:00 PM
Intercontinental Los Angeles Downtown, Floor 5 : Silver Lake A

Chair:
Yannick Nsani (Morgan State University)

Discussant:
Hyeri Hong (California State University, Fresno)

Presentations:

1. **Supporting Assessment System Coherence Through Formative Alignment Evaluation**
Brooke Nash, University of Kansas; Meagan Karvonen, ATLAS, University of Kansas ; Russell Swinburne Romine, University of Kansas

A comprehensive alignment framework centering on tailored alignment designs with formative and evaluative evidence to promote assessment-system coherence was used to develop an alignment evaluation plan for a multidimensional science assessment. We demonstrate how the framework informed the plan including a customized alignment study conducted as one source of evidence.

2. **A Holistic Model for Developing an Early Childhood English Language Proficiency Assessment**
Yu-Lan Su; Aubrey Sahouria, Center for Applied Linguistics; Jasmine Tsai, Center for Applied Linguistics; Yamei Wang, Center for Applied Linguistics

This two-year study outlines the development of a psychometrically sound English language proficiency assessment for early childhood learners, leveraging storybook-based tasks to enhance engagement. The assessment underwent a rigorous multi-phase process to achieve strong psychometric characteristics. This presentation details the methodological framework and analyses that underpin the final assessment forms.

3. **Developing a four-dimensional Computational Thinking Assessment Framework and Tool for STEM Readiness**
Yunting Liu, University of California, Berkeley; Linwei Yu, The University of Hong Kong; Mark Wilson, University of California, Berkeley; Karen Draney

Computational Thinking (CT) is considered a fundamental ability in STEM disciplines, and an integral part of college readiness. We present a CT framework incorporating four dimensions and an instrument created using the BEAR Assessment System. Reliability and validity evidence were gathered, helping to understand the internal structure of Computational Thinking.

4. **Examining Process-Specific Engagement and Comprehension in digitalPIRLS 2021**
   Jeneve Swaby, Boston College; Matthias von Davier, Boston College

   This study uses digitalPIRLS 2021 log data to examine time on task, comprehension processes, and accuracy in digital reading. Linear and logistic mixed-effects models were used to investigate whether engagement varies across processes, how comprehension process difficulty is reflected in accuracy, and the relationship between response time and accuracy.

5. **An application of justice-oriented instrument development and validation**
   Megan Welsh; Valeria Zunino, University of California, Davis; Matthew Wallace, University of California, Davis; Margarita Jimenez-Silva, University of California - Davis; Robin Martin, University of California, Davis; Tony Albano, University of California - Davis

   We examine the contributions made by teacher-collaborators from marginalized groups during development of a measure of equitable English Language Development in mathematics instruction. Analyzing transcripts and work products from collaborative design sessions, we explore their contributions, highlighting how early involvement strengthens the authenticity and justice-orientation of affective measures.

Psychometric Issues in Test Development
Individual Paper Session
3:30 PM – 5:00 PM
Intercontinental Los Angeles Downtown, Floor 7: Hollywood Ballroom II
:
Chair:
Edynn Sato (Sato Education Consulting LLC)

Discussant:
Edynn Sato (Sato Education Consulting LLC)

Presentations:

1. **Construct Waypoints: Evaluating Banding Alignment in the New DRDP'25 Assessment**
   Joshua Sussman, University of California, Berkeley

   This paper develops and applies quantitative metrics for evaluating the quality of construct waypoints—reporting bands in the new DRDP'25 assessment—grounded in developmental theory and Rasch modeling. We introduce straightforward indices of "banding congruence," test the clustering hypothesis of threshold alignment, and quantify alignment between band location and observed responses.

2. **Evaluation of Scoring Models for Innovative Multi-Part Items**
   Hannah Lewis, University of North Carolina at Chapel Hill; Francis O'Donnell, National Board of Medical Examiners; Christopher Runyon, National Board of Medical Examiners

   Multi-part items are sometimes needed to assess complex constructs, but present scoring challenges. We demonstrate a process for addressing these challenges by investigating the relationship among components of an innovative multi-part item and comparing scoring models based on model-data fit, information, ease of implementation, and implications for reporting results.

3. **Impact of Linguistic Features of Multiple-Choice Item Options on Item Functioning**
   Euigyum Kim, Sogang University; Hyo Jeong Shin, Sogang University

   This study examines how linguistic features of multiple-choice options influence item functioning. Findings reveal that lower lexical overlap with the passage, higher lexical diversity, and lower readability increase item difficulty. Greater overlap and higher readability enhance discrimination. Our study provides practical guidance for incorporating linguistic features in multiple-choice item development.

4. **The Impact of the G-DINA Model Discrimination Index Configurations on Classification Accuracy**
Zechu Feng, The University of Hong Kong; Jimmy de la Torre, The University of Hong Kong

Cognitive diagnosis models have been used in educational, psychological, and clinical assessments. In real-life applications, where test lengths are typically limited, optimal selection of items becomes crucial. This study investigates the property of the GDI and how its configuration is related to the pattern classification accuracy.

5. **A Multi-Vantage Point Perspective to Internal-Structure Validity: A Case for ROAR-Inference**
Alexander Blum, Stanford University; Robin Irey, University of California, San Francisco; Tonya Murray, Stanford Graduate School of Education; Yukie Toyama; Jason Yeatman, Stanford University; Rebecca Silverman, Stanford University

An essential ingredient to a multi-vantage point perspective to internal structure validity are sense making tools such as construct mapping and waypoints. Through Rasch family models a correlational and means perspective were taken together using ROAR Inference, used by 1171 students grades 2-5, as a case example.

Reimagining Measurement: History, Method, and Consequences across Higher Education and Statewide Testing (GSIC Session)

Coordinated Paper Session
3:30 PM – 5:00 PM
Intercontinental Los Angeles Downtown, Floor 5 : Silver Lake B

Educational measurement is often presented as a neutral technical craft, yet its histories, practices, and impacts are deeply tied to deficit logics and institutional mechanisms that reproduce inequality. This coordinated session develops a complete argument that moves from historical critique to methodological reconstruction to empirical applications in higher education and statewide testing.

Paper 1 excavates the deficit lineage of the field, showing how eugenic and racialized logics shaped modern measurement and continue to influence construct definitions and validation practices. Paper 2 codifies justice-oriented methodological standards, including participatory design, consequential validity, subgroup-specific reporting, and epistemic equity audits, offering a blueprint for equity-centered practice. Paper 3 uses narrative inquiry with Nepali Dalit students in U.S. higher education to reveal how conventional metrics misrecognize structurally excluded learners and to clarify what equitable measures must capture. Paper 4 applies psychometric modeling to technology-enhanced items in Iowa statewide assessments, testing how structural inequities across districts shape item functioning and student outcomes.

Together, the four papers demonstrate that validity is not only a question of internal coherence but also a public obligation. The session invites NCME to advance measurement practices that minimize harm and strengthen institutional accountability through design, analysis, and governance.

Chair:
Khem Sedhai

Discussant:
Kayla Burt (NCME GSIC Co-chair)

Presentations:

1. **Beyond the Metrics: A Narrative of Dalit Education and Mobility.**
   Khem Sedhai, University at Albany, SUNY

2. **Critical Quantitative Methods for Measurement Reform: Rethinking Validity, Power, and Purpose**
   Catherina Villafuerte, University of Connecticut

3. **A Historical Account of Testing and Deficit Narratives**
   Brein Mosely, Harvard University

4. **Evaluating Technology-Enhanced Items: A Critical Psychometric Analysis of Equity**
   Alexis Oakley

Research on Through-Year Assessment

Individual Paper Session
3:30 PM – 5:00 PM
Intercontinental Los Angeles Downtown, Floor 7: Roosevelt B

Chair:
Merve Sarac (College Board)

Discussant:
Laurie Davis (Curriculum Associates)

Presentations:

1. **Impact of Routing Decisions on End-of-year Student Outcomes in Through-Year Multistage Test**
   Sangdon Lim, Cambium Assessment

   In a through-year multistage test, students who have the same true ability can be routed to different pathways, which may lead to undesirably different end-of-year outcomes. Using simulated data from a six-stage through-year assessment, the impact of routing decisions on end-of-year student ability estimates was examined using regression discontinuity analyses.

2. **Comparison of Through-Year Multistage Assessments with Different Module Lengths in Early Seasons**
   Sangdon Lim, Cambium Assessment; Christina Schneider, Cambium Assessment; Garron Gianopulos, Cambium Assessment

   When progress monitoring is the goal, stakeholders often desire fewer items in the fall and winter compared to the spring. Using simulated data from a six-stage assessment, we examined how this affects ability estimates and routing accuracy, which are key factors for growth measures and accountability systems.

3. **A Pragmatic Methodology for Validating a State's Through-Year Assessment**
   Jonathan Downey, New Meridian Corporation / Carnegie Mellon University Enhance Program; Christopher Gentile, New Meridian Corporation; Cedar Rose, Montana Office of Public Instruction

   A validation study of a state's through-year assessment suggested that students prefer the format over traditional alternatives and that professional development is key for teachers using score reports effectively. However, generating clear and actionable findings required creative and pragmatic methodological strategies, offering lessons for other states validating innovative assessment systems.

4. **Aggregating Instructionally Embedded Assessment Data Using Beta IRT Models**
   Auburn Jimenez, University of Kansas; Jake Thompson, ATLAS, University of Kansas; Brooke Nash, University of Kansas

   Beta IRT models are a psychometric framework that can integrate mastery-based reporting with scale score-based reporting, supporting outcomes that balance instructional and policy needs. In the present study, we demonstrate the utility of Beta IRT models using Pathways for Instructionally Embedded Assessment (PIE) as a case study.

5. **Evaluating the 2024–2025 Texas Through-Year Assessment Pilot (Year 3) Participation Efficacy**
   Yuanyuan McBride, Pearson; Michael Chajewski, Pearson

   This study continued the 2023–2024 evaluation of TTAP by analyzing matched TTAP and non-TTAP examinees across multiple 2024–2025 STAAR subjects. Using Coarsened Exact Matching, weighting, and regression analyses, participation efficacy varied across subjects, and TTAP scale scores were statistically significant predictors of STAAR performance with measurable effect sizes.

Sources of Contemporary and Remaining Challenges in Large Scale NGSS Assessment Development

Organized Discussion
3:30 PM – 5:00 PM
Intercontinental Los Angeles Downtown, Floor 5 : Ladera Heights

While not as frequently tested as Mathematics and English Language Arts, science also has 21st century NextGen standards that have been adopted or adapted by a majority of states. However, the Next Generation Science Standards have proven exceedingly difficult to develop well-aligned assessments for, as predicted virtually upon their original publications—only in part because of the challenges of the NGSS construct itself.

Most of these challenges stem from differences in priorities and expectations—and perhaps even values—held by different key stakeholder roles in the assessment development process. Achieve, an early leader of the NGSS project, has long since acknowledged these different stakeholders and their different responsibilities to science education, but it has not acknowledged the tensions that exist between them.

This panel brings together a variety of professional science assessment stakeholders (i.e., content development professionals, a psychometrician and a state assessment director) to consider the contemporary and remaining challenges to the development of well-aligned and appropriate large scale NGSS assessment. (Most panelists have classroom experience, so that perspective will also inform the discussion.) They will draw out their contrasting views of the requirements for such assessments and the constraints they view as out of their discipline's control.

Chair:
Alexander Hoffman (AleDev Research & Consulting)

Discussant:
Alexander Hoffman (AleDev Research & Consulting)

Presenter(s):
Kristen Crawford, New Hampshire Department of Education; Christy Glore, ATLAS, University of Kansas; Daisy Rutstein; Katie Schmidt, College Board; Alexander Hoffman, AleDev Research & Consulting

Text/Speech-based Approaches to Item Parameter Modeling

Coordinated Paper Session
3:30 PM – 5:00 PM
Intercontinental Los Angeles Downtown, Floor 5 : Boyle Heights

With increasing needs of adaptive assessments and personalized learning, both assessment and learning analytics fields explored the use of the latest AI technology to predict item difficulty based on item text. This session spotlights a refreshed interest in assessment: using AI to infer item and passage difficulty as well as item discrimination directly from item content (text and speech) thereby reducing reliance on costly, exposure prone field testing. Across five studies, large language models (LLMs) and modern transformers extract information from stems, options, and passages, simulate or proxy examinee behaviors, and convert pairwise judgments into calibrated difficulty scales. Multitask and long context modeling strengthen pre–field test estimates of difficulty and discrimination; pairwise scaling (BTL/Thurstone) turns LLM judgments into coherent metrics; and synthetic data generated for rationales, responses, and read aloud audio with controllable fluency/prosody enables scalable experimentation without burdensome data collection. Collectively, these approaches demonstrate the promise text/speech-based approaches to modeling item and passage psychometric quality. They also open avenues for explainability (e.g., option level signals), for integrating content aware predictions with item response theory and for advancing adaptive testing and personalized instruction. In sum, content driven, AI assisted methods can augment traditional measurement methods in item parameter modeling.

Chair:
Hotaka Maeda (Smarter Balanced)

Discussant:
Hotaka Maeda (Smarter Balanced)

Presentations:

1. **Text-based Approaches to Item Difficulty and Discrimination Modeling**
   Hong Jiao, University of Maryland; Hanna Choi, Ewha Womans University; Sydney Peters, University of Maryland, College Park; Ming Li, University of Maryland; Tianyi Zhou, University of Maryland

2. **Item Response Modeling with Textual Information**
   Andrew Lan, University of Massachusetts, Amherst; Wanyong Feng, University of Massachusetts, Amherst; Alexander Scarlatos, University of Massachusetts Amherst; Nigel Fernandez, University of Massachusetts Amherst; Peter Tran, University of Massachusetts, Amherst; Christopher Ormerod; Susan Lottridge, Pearson; Stephen Sireci, University of Massachusetts Amherst

3. **Scaling Item Difficulty Using LLM-Based Pairwise Judgments: An Application of the Bradley-Terry-Luce Model**
   Suhwa Han, Cambium Assessment Inc.; Frank Rijmen, Cambium Assessment; Christopher Ormerod; Susan Lottridge, Pearson

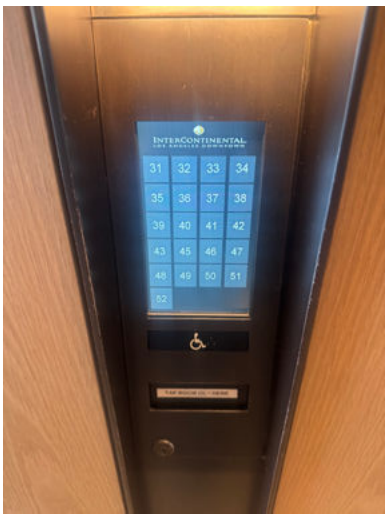
4. **Predicting Item Statistics: Thurstone Scaling with Large Language Models**
   Nate Smith, American Board of Internal Medicine; Kelly Rewley

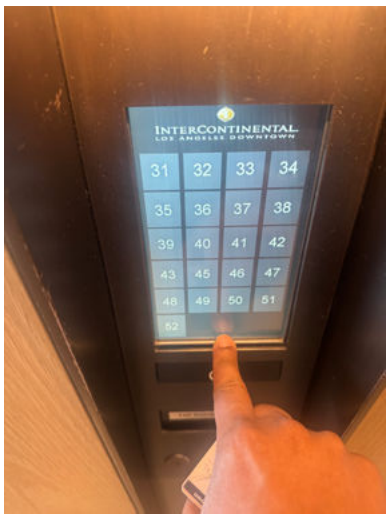
5. **Generating Synthetic Read-Aloud Audio Data for Passage Difficulty Modeling in ORF Assessment**
   Kuo Wang, Southern Methodist University; Akihito Kamata

Closing Reception
Social
5:30 PM – 7:00 PM
InterContinental at Dekkadance

To access Dekkandance, you may take the spiral staircase from the InterContinental lobby on the 70th floor -or- you may take the “C” and “D” bank elevators to the 69th floor.

Note from hotel on “C” bank elevators: it can get tricky, you will need to press something towards the bottom of the keypad in the “blank” area, then you will be able to manually input “69.” Here are photos for reference:





PARTICIPANT INDEX

(Last Name, First Name,Affiliation, Email)

., Ketan, ketan@umass.edu, University of Massachusetts, Amherst
Abroampah, Doris, abroampa@ualberta.ca, University of Alberta
Abulela, Mohammed, mhady001@umn.edu, MetaMetrics, Inc. and University of Minnesota
Acharya Julien, Nisha, nisha.acharya@gmail.com, Center for Measurement Justice
Adams, Daniel, dadams@ets.org, ETS
Adegoke, Hope, HOADEGOKE@uncg.edu, University of North Carolina, Greensboro
Albano, Tony, adalbano@ucdavis.edu, University of California - Davis
Alexander, Patricia, palexand@umd.edu, University of Maryland
Alila, Nancy, nancy.alila@uga.edu, Department of Educational Psychology, University of Georgia, Athens, GA, USA
Almond, Russell, ralmond@fsu.edu, Florida State University
Altamimi, Tasnim, tzaltamimi@uncg.edu, University of North Carolina at Greensboro
Alzabidi, Talal, tsalzabidi@uncg.edu, University of North Carolina Greensboro
Ames, Allison, aames@nbme.org,
An, Xiaozhu, xan@ixl.com, IXL Learning
Andrews, Benjamin, andrewsjbenjamin@gmail.com,
Anghel, Ella, anghel@bgu.ac.il, Ben-Gurion University in the Negev
Araneda, Sergio, sondaxius@gmail.com, Caveon Test Security
Arbet, Gregory, gaarbet@utexas.edu, University of Texas at Austin
Arias, Angel, AngelArias@cunet.carleton.ca, Carleton University
Arslan, Burcu, BARSLAN@ets.org, ETS Research Institute
Arthur, David, dbarthur@uw.edu, University of Washington
Arthur, Ann, Ann.arthur@act.org, ACT Education Corp.
Asamoah, Nana Amma, nab.asamoah@gmail.com, University of Arkansas
Asare, Eric, asareeric12@gmail.com, Old Dominion University
Asare, Eric, easar004@odu.edu, Old Dominion University
Asher, Michael, masher@andrew.cmu.edu, Carnegie Mellon University
Atasever, Umut, umut.atasever@iea-hamburg.de, IEA
Attali, Yigal, yigal@duolingo.com, Duolingo
Azim, Farhan, farhan.azim@unimelb.edu.au, The University of Melbourne
Badrinarayan, Aneesha, aneesha.badrinarayan@gmail.com, Education First
Bailey, Paul, paul.dean.bailey@gmail.com, American Institutes For Research
Balisciano, Nicholas, nbalisciano@alumni.upenn.edu, Harvard University
Ballantyne, Keira, kballantyne@cal.org, Center for Applied Linguistics
Banjanovic, Erin, erin.banjanovic@gmail.com,
Bao, Yu, bao2yx@jmu.edu, James Madison University
Baral, Kushmakar, Kushmakar.Baral@du.edu, University of Denver
Barber, Justin, justin.barber@pearson.com, Pearson
Barnes, Sophie, sophie.barnes@yale.edu, Yale University
Barnum, Stuart, stuartbarnum@gmail.com, National Board of Osteopathic Medical Examiners
Barros-Bustos, Solange, solange.barrosb@ku.edu, The University of Kansas
Beach, Sarah, seb3mr@virginia.edu, University of Virginia
Beard, Jonathan, jonathan.j.beard@gmail.com, College Board
Bediwy, Ahmed, ahmed-bediwy@uiowa.edu, The University of Iowa
Behrens, John, jbehrens@nd.edu, University of Notre Dame
Beiting-Parrish, Magdalen, magdalen.beiting@gmail.com, EdAIfy, CUNY Graduate Center
Beiting-Parrish, Magdalen, mbeiting@gradcenter.cuny.edu,
Bell, Courtney, courtney.bell@wisc.edu, University of Wisconsin - Madison
Belzak, William, wbelzak@gmail.com, Duolingo
Bennett, Randy, randyebennett@gmail.com, Assessment Innovation Matters
Betebenner, Damian, dbetebenner@nciea.org, Center for Assessment
Betts, Joe, jbetts5118@aol.com, National Council of State Boards of Nursing
Bezirhan, Ummugul, bezirhan@bc.edu, Boston College
Biancarosa, Gina, Ginab@uoregon.edu, University of Oregon
Bishop, Kyoungwon, kei.bishop2@gmail.com,
Bishop, Kyoungwon, kbishop3@wisc.edu, University of Wisconsin - Madison

PARTICIPANT INDEX

(Last Name, First Name,Affiliation, Email)

Blackmon, Elizabeth, Elizabeth.Blackmon@gcpsk12.org, Gwinnett County Public Schools
Block, Staci, slblock2021@gmail.com, California Indian Education for All
Blum, Alexander, AMBlum@EnrichYourAcademics.com, Stanford University
Bolt, Daniel, dmbolt@wisc.edu, University of Wisconsin - Madison
Bonge, Nicole, ngbonge@uark.edu, American Board of Internal Medicine
Bonifay, Wes, bonifayw@missouri.edu, University of Missouri
Botter, Preston, pdbotter@gmail.com,
Boyd, Aimee, aimeeboyd@cainc.com, Curriculum Associates
Bradshaw, Laine, lainebradshaw@gmail.com, Pearson
Brandt, Christopher, cbrandt@nciea.org, Center for Assessment
Brenchley, Mark, mark.brenchley@cambridge.org, Cambridge Assessment
Briggs, Derek, derek.briggs@colorado.edu, University of Colorado - Boulder
Brookhart, Susan, suebrookhart@gmail.com, Duquesne University
Brown, Richard, rich@westcoastanalytics.com, West Coast Analytics
Bryer, Jason, jason@bryer.org, City University of New York
Buchbinder, Nicolas, nicolas.buchbinder@colorado.edu, University of Colorado Boulder
Buckendahl, Chad, cbuckendahl@acsventures.com, ACS Ventures, LLC
Bui, Sao, thiensao.bui@student.unimelb.edu.au, The University of Melbourne
Bulut, Okan, bulut@ualberta.ca, University of Alberta
Burleigh, Tyler, tylerb@khanacademy.org, Khan Academy
Burt, Kayla, kaylabur@buffalo.edu, NCME GSIC Co-chair
Butterbaugh, Donna, donnajbb@comcast.net, ISC2
Buzick, Heather, hmmann@gmail.com, ACT
Cakici-Eser, Derya, deryacakicieser@gmail.com, Pearson
Campbell, Ian, ian.campbell@cambiumassessment.com, Cambium Assessment
Cancado, Luciana, lcancado@cainc.com, Curriculum Associates
Cao, Yichong, yichong.cao@doe.k12.de.us, Delaware Department of Education
Cardwell, Ramsey, ramsey@duolingo.com, Duolingo
Carr, Peggy, peggycarr36@yahoo.com, Former Commissioner, National Center for Education Statistics, U.S. Department of Education
Castro, Mariana, mcastro@wisc.edu, University of Wisconsin - Madison
Chakraborty, Roti, rchakraborty3@student.gsu.edu, Georgia State University, American Institutes for Research
Chao, Hsiu-Yi, psyhyc@scu.edu.tw, Soochow University
Chavez, Carlos, chave143@umn.edu,
Chen, Jyun-Hong, psyjhc@gs.ncku.edu.tw, National Cheng Kung University
Chen, Hong, hchen102@uiowa.edu, University of Iowa
Chen, Keyu, keyu.q.chen@gmail.com, NCBE
Chen, Jing, jingchen2010@gmail.com, Khan Academy
Chen, Qipeng, qipengchen_psy@outlook.com, University of Alabama
Chen, Guanyu, chenguanyu.ubc@gmail.com, The University of British Columbia
Chen, Xiuhan, xcthc@missouri.edu, University of Missouri
Chen, Siyuan, marco@duolingo.com, Duolingo, Inc.
Chen, Yiming, chen9462@umn.edu, University of Minnesota
Cheng, Ying, ycheng4@nd.edu, University of Notre Dame
Cheng, Britte, bcheng@menloedu.org, Menlo Education Research
Cheng, Yijun, chengxb@uw.edu,
Chiu, Chia-Yi, cc5010@tc.columbia.edu, Teachers College, Columbia University
Cho, Youngmi, youngmi.cho@riversideinsights.com, Riverside Insights
Choi, Hye-Jeong, hchoi@humrro.org, HumRRO
Choi, Jinah, Jinah.Choi@edmentum.com, Edmentum
Choi, Jeongwon, jeongwon.choi@vanderbilt.edu, Vanderbilt University
Choi, Mihye, mchoi@cainc.com, Curriculum Associates
Christensen, Wendy, wchriste@gmail.com, University of Colorado School of Medicine
Christensen, Laurene, llchristens2@wisc.edu, WIDA at the University of Wisconsin-Madison
Christiansen, Andrés, andres.christiansen@iea-hamburg.de, IEA Hamburg

PARTICIPANT INDEX

(Last Name, First Name,Affiliation, Email)

Chung, Jinmin, jinmin-chung@uiowa.edu, Univ. of Iowa
Chung, Bomin, schatz030714@gmail.com,
Chung, Hyewon, hyewonchung7@gmail.com, Chungnam National University
Cintron, Dakota, dwcintron@gmail.com, Claremont Graduate University
Cipriano, Christina, christina.cipriano@yale.edu, Yale University
Clark, Lisa, Lisaclark5077@gmail.com, City University of New York
Cochron, Mary, mary@metimured.com, Metimur Educational Measurement
Cordoba Perozo, Michel, mcordob@purdue.edu, Purdue University
Crespo Cruz, Eduardo J, ecrespocruz@umass.edu, University of Massachusetts, Amherst
Crosby, Michelle, michelle@amstat.org, American Statistical Association
Cruz, Sandra, spcruz@gwu.edu, George Washington University
Cui, Wenju, wcui@ets.org, ETS
Dadey, Nathan, ndadey@nciea.org, The National Center for the Improvement of Educational Assessment
Dadzie, Justice, jdadzie@crimson.ua.edu, The University of Alabama
Daisher, Ted, t3daisher@gmail.com,
Davidson, Anne, annie@crescendoed.net, CrescendoEd LLC
Davis, Laurie, laurie@davistx.com, Curriculum Associates
Davis-Becker, Susan, sdavisbecker@acsventures.com, ACS Ventures, LLC
D'Brot, Juan, jdbrot@nciea.org, Center for Assessment
De Boeck, Paul, deboeck.2@osu.edu, The Ohio State University
de la Torre, Jimmy, j.delatorre@hku.hk, The University of Hong Kong
de Leon, Margaret, margaretcdeleon@gmail.com, University of Toronto
de los Reyes, Andy, adlr@umd.edu,
Demirkaya, Onur, onurdmrkaya@gmail.com, Riverside Insights
Depré, Katharina, kdepre@uni-mainz.de, Johannes Gutenberg University Mainz
Deters, Lauren, lauren.fluegge@gmail.com, Khan Academy
Deters, Lauren, laurendeters@khanacademy.org, Khan Academy
Dicerbo, Kristen, kdicerbo@cox.net, Khan Academy
Dilek, Ismail, ismaildilek88@gmail.com,
DiStefano, Christine, distefan@mailbox.sc.edu, University of South Carolina
Domaleski, Christopher, cdomaleski@nciea.org, Center for Assessment
Dong, Yaxin, yd19@ualberta.ca, university of alberta
Donoghue, John, jdonoghue@ets.org, ets
Doran, Harold, hdoran@humrro.org, HumRRO
Douglas, Aaro, adouglas@newmeridian.org, New Meridian Corporation
Downey, Jonathan, jdowney@newmeridian.org, New Meridian Corporation / Carnegie Mellon University Enhance Program
Du, Ying, ydu@abpeds.org, American Board of Pediatrics
Du, Kyle, kdu@gradcenter.cuny.edu, CUNY Graduate Center
Dugdale, Debbie, debbie.dugdale@cambiumassessment.com, Cambium Assessment, Inc.
Dunbar, Stephen, steve-dunbar@uiowa.edu, University of Iowa
Dunn, Jennifer, jdunn@collegeboard.org, The College Board
Duran, Lillian, lduran@uoregon.edu, University of Oregon
Eckerly, Carol, caroleckerly@gmail.com, ABIM
Egan, Karla, karla.egan@edmetric.com, EdMetric
Embretson, Susan, susan.embretson@psych.gatech.edu, Georgia Institute of Technology
Englert, Kerry, kenglert@senecaconsulting.org, Seneca Consulting, LLC
Ercikan, Kadriye, kercikan@ets.org, ETS Research Institute
Everson, Howard, howard.everson@gmail.com, City University of New York
Fan, Fen, ffan@nbme.org, National Board of Medical Examiners
Fan, Meng, mfan@humrro.org, HUMRRO
Federiakin, Denis, denis.federiakin@gmail.com, Johannes Gutenberg University of Mainz; Goethe University of Frankfurt
Feinberg, Richard, RFeinberg@nbme.org, NBME
Feuerstahler, Leah, lfeuerstahler@fordham.edu, Fordham University
Filonczuk, Audrey, afiloncz@nd.edu,

PARTICIPANT INDEX

(Last Name, First Name,Affiliation, Email)

Fina, Anthony, anthony-fina@uiowa.edu, University of Iowa
Fincher, Melissa, mfincher@edcount.com, edCount, LLC
Finnegan, Robert, rfinnegan@ets.org, ETS
Finney, Sara, finneysj@jmu.edu, James Madison University
Fishbein, Bethany, fishbeib@bc.edu, Boston College
Fisk, Charles, fisk@umd.edu, NBOME
Forsyth, Carol, cforsyth@ets.org, ETS
Forte, Ellen, eforte@edcount.com, edCount, LLC
Foster, Paul, pdfoster@smu.edu, Southern Methodist University
Foster, David, david.foster@caveon.com,
Frey, Sharon, sharon.frey@riversideinsights.com, Riverside Insights
Frohn, Scott, scottfrohn@gmail.com, Khan Academy
Fu, Yanbin, ybfu@umd.edu, University of Maryland, College Park
Fukuhara, Hirotaka, hiro.fukuhara@pearson.com, Pearson
Furgol Castellano, Katherine, KEcastellano@ets.org, ETS Research Institute
Gao, Yizhu, Yizhu.Gao@uga.edu, University of Georgia
Gardner, Tracy, tracy.gardner1@pearson.com, Pearson
Gazit, Noa, noag@nite.org.il,
Geisinger, Kurt, kgeisinger2@unl.edu,
Gholson, Melissa, melissa.gholson@gmail.com, ATLAS, University of Kansas
Gianopulos, Garron, garron.gianopulos@cambiumassessment.com, Cambium Assessment
Gilbert, Joshua, Joshua_gilbert@gse.harvard.edu,
Gitiria, Lucy, lgitiria@umass.edu, University of Massachusetts, Amherst
Golden, Richard, golden@utdallas.edu, University of Texas at Dallas
Gonzalez, Jorge, jorge.gonzalezb@uc.cl, Pontificia Universidad Catolica de Chile
Gooch, Reginald, rmgooch@ets.org, ETS
Goodman, Joshua, joshuag@nccpa.net, National Commission on Certification of Physician Assistants
Gorgun, Guher, gorgun@ualberta.ca, Leibniz Institute for Science and Mathematics Education
Gorney, Kylie, kgorney@msu.edu, Michigan State University
Grabovsky, Irina, igrabovsky@nbme.org,
Graesser, Arthur, graesser@memphis.edu, University of Memphis
Grochowalski, Joe, jgrochowalski@collegeboard.org, College Board
Gui, Yi, queenmq@gmail.com, Measurement Incorporated (MI)
Gundula, Archangel, agundula2@huskers.unl.edu, University of Nebraska-Lincoln (UNL)
Gunes, Sena, gunesse@mef.edu.tr, MEF University
Guo, Hongwen, hguo@ets.org, ETS Research Institute
Guo, Wenjing, wenjing.guo@pearson.com, Pearson
Guo, Yage, yageguo@gmail.com,
Guo, Jingyi, jguo9@nd.edu, University of Notre Dame
Ham, Eun Hye, thanks02@gmail.com, Kongju National University
Hamilton, Laura, lhamilton@nciea.org, National Center for the Improvement of Educational Assessment, Inc.
Hampel, Jean, jeanhampel9@gmail.com, HMH
Han, Kyung (Chris), truetheta@gmail.com, GMAC
Han, Suhwa, suhwa@utexas.edu, Cambium Asessment
Han, Suhwa, Suhwa.Han@cambiumassessment.com, Cambium Assessment Inc.
Han, Kahee, kaheehan@ku.edu, University of Kansas
Hansen, Mark, markhansen@ucla.edu,
Hao, Jiangang, jhao@ets.org, ETS
Hardy, Michael, hardym@stanford.edu, Stanford University
Harra, Kjorte, kjorteh@gmail.com, University of Wisconsin-Madison
Harrold, Brian, bharrold@email.sc.edu, University of South Carolina
Hatch, Sarah, shatch@g.harvard.edu, Harvard University
Haviland, Sara, shaviland@ets.org, ETS
Hayes, Heather, heatherschema@gmail.com, Western Governors University
He, Qiwei, qiwei.he@georgetown.edu, Georgetown University

PARTICIPANT INDEX

(Last Name, First Name,Affiliation, Email)

He, Yi, yi.he@edmentum.com, Edmentum
He, Surina, surina1@ualberta.ca, University of Alberta
Hebert, Andrea, andreahebert66@gmail.com, Cognia
Hebert, Cristyne, cristyne.hebert@uregina.ca, Associate Professor
Hemenway, Michael, michael.hemenway@pearson.com, Pearson
Hendrickson, Amy, ahendrickson@collegeboard.org, The College Board
Henriques, Jonathan, henri2jm@dukes.jmu.edu, James Madison University
Hernández, Diley, diley.hernandez@gatech.edu, Georgia Institute of Technology
Herr, Riley, herrrk@jmu.edu, James Madison University
Hibbard, Susan, susan.hibbard@theaba.org, The American Board of Anesthesiology
Hilliard, Paul, philliard@ets.org,
Himelfarb, Igor, ihimelfarb@nbce.org,
Ho, Andrew, Andrew_Ho@gse.harvard.edu, Harvard University
Ho, TsungHan, tho@ets.org, ETS
Hoffman, Alexander, ahoffman@AleDev.com, AleDev Research & Consulting
Holliday, Keith, kholliday@cainc.com, Curriculum Associates
Holtmann, Marlen, holtmann.marlen@gmail.com, IPN - Leibniz Institute for Science and Mathematics Education, EUF – Europa-Universität Flensburg
Hong, Hyeri, hyerihong@mail.fresnostate.edu, California State University, Fresno
Hong, Minju, minjuh1215@cau.ac.kr, Chung-Ang University
Hong, Youmin, ymhong@umd.edu, University of Maryland
Hong, Seong Eun, shong@cal.org, Center for Applied Linguistics
Hua, Cheng, chua@montevallo.edu, University of Montevallo
Huan, Yingqi, yh3755@tc.columbia.edu, Teachers College, Columbia University
Huang, Yue, yueh@udel.edu, Measurement Incorporated
Huang, Yingshi, yingshi@ucla.edu, University of California - Los Angeles
Huang, Jing, jing.huang1222@gmail.com, Purdue University
Huang, Wei, whuang20@crimson.ua.edu, Universtiy of Alabama, College of Education
Huang, Qi, huan2304@purdue.edu, Purdue University
Huff, Kristen, khuff@cainc.com, Curriculum Associates
Hughes, Gerunda, gerunda.hughes@gmail.com,
Hunsberger, Josiah, josiahahunsberger@gmail.com, James Madison University
Illmann, Jannick, j.illmann@dipf.de, DIPF | Leibniz Institute for Research and Information in Education
Immanuel, Aria, aimmanuel@umass.edu, University of Massachusetts Amherst
Ingrisone, James, james.ingrisone@pearson.com, Pearson VUE
Ingrisone, Soo, singrisone@gmail.com, HumRRO
Issayeva, Laila, issayelx@dukes.jmu.edu, James Madison University
Jacobs, Gregory, greg.jacobs@pearson.com, Pearson
JAFFARI, FATHIMA, romisa.2012sa@gmail.com, PSYCHOMETRIC EXPERT AT ETEC QIYAS
Jeong, Tai Sun, taisunjeong@gmail.com, University of Wisconsin-Madison
Jerez, Daniel, jerezgar@ualberta.ca, University of Alberta
Jewsbury, Paul, pjewsbury@ets.org, ETS
Jiang, Yang, yjiang002@ets.org, ETS
Jiao, Hong, hjiao@umd.edu, University of Maryland
Jimenez, Auburn, auburn.jimenez@ku.edu, University of Kansas
Jin, Kuan-Yu, kyjin@hkeaa.edu.hk, Hong Kong Examinations and Assessment Authority
Johnson, Evelyn, ejohnson@riversideinsights.com, Riverside Insights
Johnson, Matthew, msjohnson@ets.org, ETS Research Institute
Jones, Andrew, ajones@absurgery.org, American Board of Surgery
Jong, Jae Jun, jaejunj2@illinois.edu,
Jonson, Jessica, jjonson2@unl.edu, Buros Center for Testing - University of Nebraska Lincoln
Joo, Sean, sjoo@ku.edu, University of Kansas
Jung, Ji Yoon, jiyoon.jung@bc.edu, Boston College
Jung, Ae Kyong, aekyong-jung@uiowa.edu, University of Iowa
Jung, Juyoung, juyoung-jung@uiowa.edu, University of Iowa

PARTICIPANT INDEX

(Last Name, First Name,Affiliation, Email)

Jung, HyunJoo, hjung1@inha.edu, Inha University
Jurich, Daniel, DJurich@nbme.org, National Board of Medical Examiners
Kaçak, Tugay, kacaktugay@gmail.com, Trakya University
Kahne, Joseph, jkahne@ucr.edu, University of California, Riverside
Kamei, Toshiko, kamei.t@unimelb.edu.au, University of Melbourne
Kang, Yoonjeong, yoonjeong.kang@cambiumassessment.com, Cambium Assessment
Kanopka, Klint, klint.kanopka@nyu.edu, New York University
Kapoor, Shalini, shalinikapoor.ia@gmail.com,
Kara, Yusuf, ykara@miami.edu, University of Miami
Karaer, Gamze, gamze.karaer@wsu.edu, Washington State University
Karamese, Hacer, karamese@wisc.edu, WIDA at the University of Wisconsin-Madison
Karvonen, Meagan, karvonen@ku.edu, ATLAS, University of Kansas
Kataoka, Leyna, lkataoka@ucdavis.edu, University of California, Davis
Kehat, Gitit, gitit.kehat@cambiumassessment.com, Cambium Assessment
Kehinde, Olasunkanmi, ojkehinde@nsu.edu, Norfolk State University
Keller, Lisa, lkeller@umass.edu, University of Massachusetts - Amherst
Kennedy, Patrick, ppaine@uoregon.edu, University of Oregon
Kern, Justin, kern4@illinois.edu, University of Illinois at Urbana-Champaign
Keum, Eunhee, keum@cresst.org, ELPA21 at UCLA CRESST
Kevelson, Marisol, mkevelson@ets.org, ETS
Kilenthong, Weerachart, tee@riped.utcc.ac.th, University of the Thai Chamber of Commerce (UTCC)
Kim, JongPil, jp.kim@riversideinsights.com, Riverside Insights
Kim, Stella, stella-kim@charlotte.edu, University of North Carolina Charlotte
Kim, Sunhee, sunnyacct8@gmail.com,
Kim, YoungKoung, ykim@collegeboard.org, College Board
Kim, Hahyeong, hk33@illinois.edu, University of Illinois at Urbana-Champaign
Kim, Sohee, skim@southalabama.edu, University of South Alabama
Kim, Hyunah, hyunah.kim@eqao.com, Education Quality and Accountability Office
Kim, Nana, nkim530@umn.edu, University of Minnesota
Kim, Yejin, dpwls3360@snu.ac.kr, Seoul National University
Kim, Euigyum, euigyum.kim@gmail.com, Sogang University
Kim, Minjung, j145950@naver.com, Konkuk University
Kim, Yonggi, yonggi1124@gmail.com, Chungbuk National University
Kim, Yoojoong, yk96666@uga.edu, University of Georgia
Kim, Seongeun, seongeunkim8@gmail.com, National Commission on Certification of Physician Assistants
Kimble, Yasmene, kimbleyasmene@gmail.com, Stony Brook University
Kincaid, Heath, hkincaid@abog.org, American Board of Obstetrics and Gynecology
Kintz, Tara, kintztar@msu.edu, Michigan Assessment Consortium

PARTICIPANT INDEX

(Last Name, First Name,Affiliation, Email)

Klesch, Heather, Heather.klesch@pearson.com, Pearson
Kollias, Charalambos (Harry), ckollias@polytomous.com, Polytomous Limited
Kong, Lingchen, l.kong@ufl.edu, University of Florida
Koo, Jin, joykoo2000@gmail.com, Enrollment Management Association
Koo, Miryeong, koo21@illinois.edu, University of Illinois at Urbana-Champaign
Koval, Jayma, jayma.koval@ceismc.gatech.edu, Georgia Institute of Technology
Krist, Andrew, atkrist@crimson.ua.edu,
Krost, Kevin, kevinkrost@vt.edu,
Krumm, Andrew, aekrumm@umich.edu,
Kuang, Huan, hkuang2@fsu.edu, Florida State University
Kukea Shultz, Pohai, pohai@hawaii.edu, University of Hawaii - Manoa
Kumar, Lavanya Shravan, lkumar@gmac.com, Graduate Management Admission Council
Kwako, Alexander, Alexander.Kwako@cambiumassessment.com, Cambium Assessment
Kwako, William, wkwakow@gmail.com, Georgia Institute of Technology
Kwon, Sunbeom, sunbeom2@illinois.edu, University of Illinois, Urbana-Champaign
Kyllonen, Patrick, pkyllonen@ets.org, ETS
Kyung, Jungwon, jkyung@umass.edu, University of Massachusetts, Amherst
Lahoud, Tamlyn, tamlyn.lahoud@uga.edu, University of Georgia
Lakin, Joni, Jlakin@ua.edu, University of Alabama
Lambert, Laura, laycocla@jmu.edu, James Madison University
Lan, Andrew, andrewlan@cs.umass.edu, University of Massachusetts, Amherst
Lane, John, lanejoh3@msu.edu, Michigan Assessment Consortium
Lavigne, Cheryl, clavigne@ets.org, ETS
Le, Trung, trungle467@gmail.com, University of Illinois Urbana-Champaign
Lee, Eunji, ejlee@uga.edu, University of Georgia
Lee, Youngjun, youngjun.lee@theaba.org, The American Board of Anesthesiology
Lee, Haneul, haneullee@umass.edu, University of Massachusetts Amherst
Lee, Dukjae, gxe5wh@virginia.edu, University of Virginia
Lee, Won-Chan, won-chan-lee@uiowa.edu, University of Iowa
Lee, Haeju, hlee@uncg.edu, University of North Carolina Greensboro
Lee, Hyeryung, hyeryung.lee@okstate.edu, Oklahoma State University
Lee, Sunhyoung, hongasunny@gmail.com, Ascend Learning/ University of Nebraska-Lincoln
Lee, Minho, leemino72@ucla.edu, University of Notre Dame
Lee, Mina, mina.lee@cambiumassessment.com, Cambium Assessment
Lee, Sooyong, slee2462@wisc.edu, WIDA at the University of Wisconsin-Madison
Lee, Chansoon (Danielle), dchanslee@gmail.com, American Board of Internal Medicine
Lee, Sung-Hyuck, sulee@gmac.com, GMAC
Lee, Mina, mina.mh.lee@gmail.com, Cambium Assessment

Lee, Levone, levone.lee@uky.edu, University of Kentucky
Leng, Dihao, dihaoleng@outlook.com, East China Normal University
Lent, Sarah, sdlent@wisc.edu, University of Wisconsin - Madison
Leventhal, Brian, leventbc@jmu.edu, James Madison University
Lewis, Hannah, hklewis@unc.edu, University of North Carolina at Chapel Hill
Li, Feifei, fli@ets.org, Educational Testing Service
Li, Dongmei, dongmei.li@act.org,
Li, Zhifei, zhf@uiowa.edu,
Li, Weiran, weiran.li@ubc.ca, The University of British Columbia
Li, Lanrong, jessicalilr2011@gmail.com, MetaMetrics, Inc.
Li, Warren, warren@ld-insights.com, Associate Research Scientist
Li, Jujia, jli183@crimson.ua.edu, University of Alabama
Li, Wenshuo, wenshuo.li@mail.mcgill.ca, McGill University
Li, Meng-Lin, ji3edki3edcd@gmail.com,
Li, Yuxuan, yl5655@tc.columbia.edu, Teachers College, Columbia University
Li, Shirley, shirley.li@savvas.com, Savvas Learning Company
Li, Jingyang, Jingyang.Li@uga.edu, University of Georgia
Liang, Min, min-liang-1@uiowa.edu, National Board of Osteopathic Medical Examiners (NBOME)
Liang, Li, llcfy2022@gmail.com,
Liao, Manqian, mancy@duolingo.com, Duolingo
Liao, Xiangyi, xy.liao@ubc.ca, University of British Columbia
Liaw, Yuan-Ling, yuan-ling.liaw@iea-hamburg.de, IEA Hamburg
Light, Erica, elight@umass.edu, UMass Amherst
Lim, Youn Seon, limyo@ucmail.uc.edu, University of Cincinnati
Lim, Hwanggyu, hglim83@gmail.com, Inha University
Lim, Sangdon, stdevlimit@gmail.com, Cambium Assessment
Lin, Pei-Ying, pei-ying.lin@usask.ca, University of Saskatchewan
Lindner, Marlit, mlindner@leibniz-ipn.de, IPN - Leibniz Institute for Science and Mathematics Education
Ling, Guangming, gling@ets.org, ETS
Lipnevich, Anastasiya, a.lipnevich@gmail.com, Queens College and the Graduate Center, City University of New York
Litschwartz, Sophie, sophie.litschwartz@mdrc.org, MDRC
Liu, Chunyan, cliu@nbme.org, National Board of Medical Examiners
Liu, Jinghua, jinghuawalkerliu@gmail.com, The National Board of Osteopathic Medical Examiners
Liu, Huan, huan.liu@riversideinsights.com, Riverside Insights
Liu, Xiaoxiao, xiaoxia6@ualberta.ca, University of Alberta
Liu, Yiqing, yiqingl@stanford.edu,
Liu, Ou Lydia, lliu@ets.org, ETS
Liu, Yunting, yunting99@berkeley.edu, University of California, Berkeley

PARTICIPANT INDEX
(Last Name, First Name,Affiliation, Email)

Liu, Guangyun, GLiu@nbome.org, NBOME
Liu, Liyuan, liyuan.liu1@pearson.com, Pearson
Liu, Lucia, lucy.xin.liu@gmail.com, Ascend Learning
Liu, Yuming, yuming.liu@cambiumassessment.com, Cambium Assessment Inc
Liu, Joyce Xinle, xinle4@ualberta.ca, University of Alberta
Liu, Qiao, qliu14@charlotte.edu, UNC Charlotte
Liu, Kaijie, liu6kx@dukes.jmu.edu, James Madision University
Long, Yunyi, ylong@nbome.org, National Board of Osteopathic Medical Examiners
Lorie, Will, wil___@gmail.com, Center for Assessment
Lottridge, Susan, susan.lottridge@cambiumassessment.com, Pearson
Lu, Ru, rlu@ets.org, Educational Testing Service
Lu, Max, maxlu@fas.harvard.edu, Harvard University
Lu, Yi-Chen, h126027278@gmail.com,
Ludovica, De Carolis, l.decarolis@campus.unimib.it, University of Milano-Bicocca
Luo, Jinwen, jevan.luo@gmail.com, UCLA
Lyons, Susan, susan@lyonsassessment.com, Lyons Assessment Consulting
Lyons, Susan, susan@lyonsassessmentconsulting.com, Lyons Assessment Consulting
Lyu, Weicong, weiconglyu@um.edu.mo, University of Macau
Lyu, Meng, meng.lyu@bc.edu,
Ma, Jing, majing@uiowa.edu, University of Iowa
Ma, Wanjing (Anya), wanjingm@stanford.edu, Stanford University
Ma, Kiya, kiya-ma@ku.edu, University of Kansas
Ma, Wenchao, wma@umn.edu, University of Minnesota
Ma, Ye, xxmlearn@gmail.com, Amazon Web Services
Ma, Ye, cherylyema@gmail.com, Amazon Web Services
Madden, Bradley, bzmadden@uncg.edu, University of North Carolina at Greensboro
Madison, Matthew, mjmadison@uga.edu, University of Georgia
Maeda, Hotaka, hotaka.maeda@gmail.com, Smarter Balanced
Magabe, Donather, ddmagabe@uncg.edu, University of North Carolina at Greensboro
Makinde, Henry, hsmakinde@uncg.edu,
Maksimchuk, Mike, mikemaksimchuk@kentisd.org, Kent Intermediate School District
Man, Kaiwen, kman@ua.edu, University of Alabama
Mann, Kaylena, mannkay@bc.edu, Boston College
Mardones-Segovia, Constanza, cam04214@uga.edu, University of California, San Diego
Marigo, Alessia, amarigo@wisc.edu, Wisconsin University - Madison
Marini, Jessica, jmarini@collegeboard.org, College Board
Marion, Scott, smarion@nciea.org, Center for Assessment
Martinez, Alfonso, alfonso.martinez@fordham.edu, Fordham University

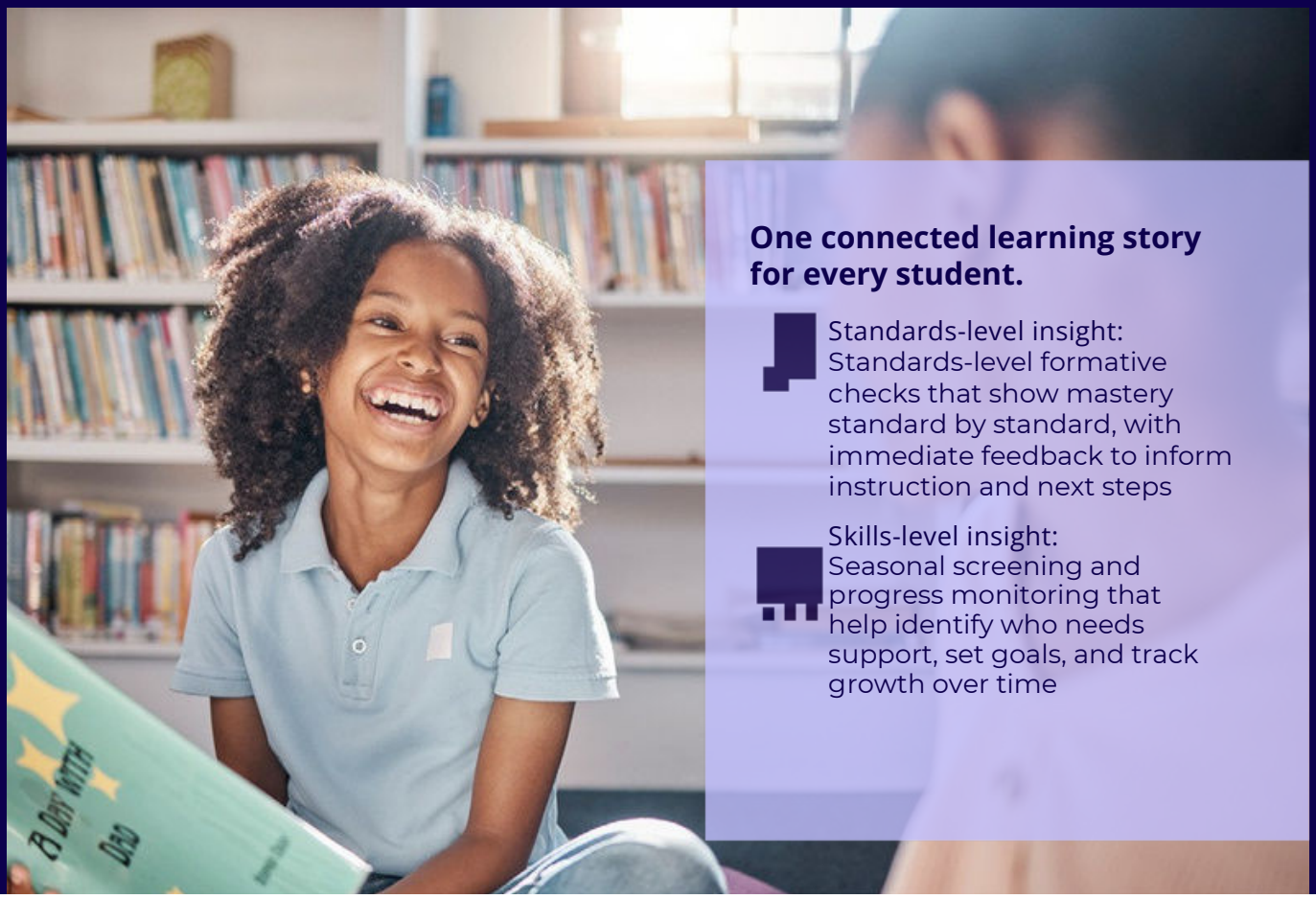
Pearson

SEE THE FULL PICTURE.
START CONNECTING THE DOTS WITH NAVVY+.

Navvy+ brings skills and standards data together in one place, so teams can move from insight to action faster.

When student data lives across systems, it's harder to act with confidence. Navvy+ unifies skills and standards data in one place, so teams can spend less time switching between systems and more time taking timely action for every learner.

Districts are working hard to give every student the right support. But when standards-level learning checks live in one place and MTSS data lives somewhere else, it's hard to see the full picture—and even harder to know what to do next.



One connected learning story for every student.

- Standards-level insight: Standards-level formative checks that show mastery standard by standard, with immediate feedback to inform instruction and next steps
- Skills-level insight: Seasonal screening and progress monitoring that help identify who needs support, set goals, and track growth over time

PARTICIPANT INDEX

(Last Name, First Name,Affiliation, Email)

Martinez, Alejandro, a123451@email.unc.edu, University of North Carolina at Chapel Hill
Martinková, Patrícia, martinkova@cs.cas.cz, Czech Academy of Sciences
Maur, Andreas, anmaur@uni-mainz.de,
Mayne, Zachary, zmayne@ixl.com, IXL
Mazzullo, Elisabetta, mazzullo@ualberta.ca, University of Alberta
McBride, Yuanyuan, malena.mcbride@pearson.com, Pearson
McCaffrey, Daniel, dmccaffrey@ets.org, ETS
McCarthy, Michael, michael.mccarthy@yale.edu, Education Collaboratory at Yale University
McClure, Kyla, Kyla.Mcclure@colorado.edu, University of Colorado Boulder
McEachin, Andrew, amceachin@ets.org, ETS Research Institute
McFadden, Mara, mcfad2me@jmu.edu, James Madison University
McMillen, Bradley, bmcmillen@wcpss.net, Wake County Public School Sustem
Mehou, Juste, mehoujm@dukes.jmu.edu, James Madison University
Mejia, Ivy, ipmejia@up.edu.ph, University of Philippines National Institute for Science and Mathematics
Mena, Fernando, fmenaserrano@umass.edu, University of Massachusetts Amherst
Meng, Huijuan, huijuam@amazon.com, Amazon Web Services
Meng, Lionel, lhmeng@wisc.edu, University of Wisconsin - Madison
Mercado, Ricardo, rmercado@datarecognitioncorp.com, DRC
Meyer, Joanna, joanna.meyer@yale.edu, The Consultation Center, Yale University
Miao, Jing, jmiao@ncsbn.org,
Michel, Rochelle, rochelle.michel@gmail.com,
Middlestead, Andrew, MiddlesteadA@michigan.gov, Michigan Department of Education
Mikeska, Jamie, jmikeska@ets.org, ETS
Mireles, Nicolas, nicolasm@mirelesconsult.com,
Mo, Ya, mooyaacn@gmail.com,
Mohamed, Moses, mosesmohamed@usf.edu, University of South Florida
Mojoyinola, Mubarak, mubarak-mojoyinola@uiowa.edu,
Moncaleano-Wallrich, Sebastian, seb.moncaleano@gmail.com,
Monroe, Scott, smonroe@educ.umass.edu, University of Massachusetts, Amherst
Morell, Linda, lindamorell@berkeley.edu, University of California, Berkeley
Morrison, Kristin, KMorrison@cainc.com,
Mosely, Brein, breinmosely@g.harvard.edu, Harvard University
Moses, Tim, tmoses2@unl.edu, Buros Center for Testing
Moskowitz, Josh, jmoskowitz@acurityinsights.com, Acuity Insights
Moustaki, Irini, I.Moustaki@lse.ac.uk, London School of Economics and Political Science
Muntean, William, wmuntean@ncsbn.org, National Council of State Boards of Nursing
Myers, Aaron, amyers@abim.org, American Board of Internal Medicine
Myoung, Eunjung, ejmyoung@stanford.edu, Stanford University
Nadela, Savira, savira@stanford.edu, Stanford University

PARTICIPANT INDEX

(Last Name, First Name,Affiliation, Email)

Nagasawa, Mirai, mnagasawa@crimson.ua.edu,
Nagpal, Pooja, pnag0736@uni.sydney.edu.au, University of Sydney
Nam, Sungjin, namsungjinn@gmail.com, ACT, Inc
Nash, Brooke, bnash@ku.edu, University of Kansas
Nastuta, Sebastian, sebastian.nastuta@pearson.com, Pearson Education UK
Nasyrova, Anna, anasyrova@umass.edu, University of Massachusetts, Amherst
Naumann, Alexander, alexander.naumann@tu-dortmund.de,
Naveiras, Matthew, matthew.naveiras@riversideinsights.com, Riverside Insights
Nelluvelil, Jerry, jnelluvelil@fas.harvard.edu, Harvard Graduate School of Education
Nese, Joseph, jnese@uoregon.edu, University of Oregon
Niu, Luping, NEWL787@gmail.com,
Niyirinda, Theode, tniyirinda@crimson.ua.edu, University of Alabama
Noh, Hyerim, ahdclal5@naver.com,
Nordmeyer, Jon, jon.nordmeyer@wisc.edu, University of Wisconsin-Madison
Nsani, Yannick, yansa1@morgan.edu, Morgan State University
Nyamulani, Opalhawaye, onyamulani@fordham.edu, Fordham University
Nydick, Steven, swnydick@gmail.com, Duolingo
Oakley, Alexis, alexis-c-oakley@uiowa.edu,
Ober, Teresa, tober@ets.org, ETS
Ocheni, Christoper, caocheni@crimson.ua.edu, The University of Alabama, Tuscaloosa
O'Donnell, Francis, fodonnell@nbme.org, National Board of Medical Examiners
Ofori Aboah, Valerie, oforiaboah.1@osu.edu, The Ohio State University
Oh, Hyunjee, hyunjeeoh1223@gmail.com, Teachers College, Columbia University
Oh, Hyeonjoo, joannehj@gmail.com, Riverside Insights
Oh, Kyuseol, kyuseol@knu.ac.kr, Hyosung girl's high school
Oh, Seung Min, smoh2@illinois.edu, University of Illinois at Urbana-Champaign
Olea, Joemari, jo28397@my.utexas.edu, University of Texas at Austin
Omonkhodion, Comfort, comforthappiness.omonkhodion@ucf.edu, University of Central Florida
Omopekunola, Moses, omopekunolamoses@gmail.com, Center of psychometrics and measurements in education, HSE University, Moscow, Russia.
O'Neil, Timothy, t66oneil@yahoo.com, Pearson
Ormerod, Christopher, christopher.ormerod@gmail.com, Cambium Assessment
Ormerod, Christopher, christopher.ormerod@cambiumassessment.com,
Oyeniran, Daniel, dooyeniran@crimson.ua.edu, The University of Alabama
Ozcan, Meltem, ozcan@usc.edu, University of Southern California
Ozkan, Fatih, fatih_ozkan1@baylor.edu, Baylor University
Padellaro, Frank, frank.padellaro@cognia.org, Cognia
Paek, Pamela, PamelaPaek@gmail.com,
Pahlen, Shandell, shandell.pahlen@state.mn.us,

CollegeBoard

Join Us in Moving Measurement Forward

In the spirit of accelerating the field of educational measurement, we're excited to demonstrate our commitment to advancing rigorous, equitable, and innovative practices that promote learning and fair assessment.

Join us for engaging conversations and informative sessions that foster collaboration, support readiness and access, and deepen our commitment to meaningful impact.

ncme
NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION

PARTICIPANT INDEX

(Last Name, First Name,Affiliation, Email)

Palermo, Corey, cpalermo@measinc.com, Measurement Incorporated
Palma, Jose, jpalma@tamu.edu, Texas A&M University
Park, Yooyoung, pyooyoung@gmail.com, National Board of Osteopathic Medical Examiners
Park, Hyemin, hyemin.park@berkeley.edu, UC Berkeley
Park, Junhee, junhee-park@uiowa.edu, University of Iowa
Park, Seongmin, smparkedu@gmail.com, Korea University
Pastor, Dena, pastorda@jmu.edu, James Madison University
Patarapichayatham, Chalie, chalie.patara@nwea.org, HMH
Patz, Richard, rpatz@berkeley.edu, University of California, Berkeley
Patz, Rich, rich.patz@ncme.org,
Peabody, Michael, michael.peabody77@gmail.com, IXL Learning
Pedersen, Blaine, blainepedersen37@gmail.com,
Peng, Fang, pfrenee@gmail.com, NWEA
Perez, Joselyn, joselyn.perez@uconn.edu,
Pham, Duy, phamduy.edu@gmail.com, New Meridian
Pham, Duy, dpham@newmeridian.org, New Meridian Corporation
Piantaggini, Lance, lpiantaggini@umass.edu, University of Massachusetts, Amherst
Pirani, Sarah, spirani@osteopathic.org, American Osteopathic Association
Pires Gifford, Laura, laura.piresgifford@wsu.edu, Washington State University
Po Hsien, Hu, bshu0619@gmail.com, National Sun Yat-sen University
Poole, Glenn, gapoole@wisc.edu, WIDA
Powers, Sonya, sopowers@gmail.com, Edmentum
Primi, Ricardo, rprimi@mac.com, Universidade São Francisco
Qiao, Xin, xqiao@usf.edu, University of South Florida
Quan, Yale, yalequan@uw.edu, University of Washington
Quesen, Sarah, sarah.quesen@gmail.com, WestEd
Quirk, Victoria, vquirk3@illinois.edu, University of Illinois at Urbana-Champaign
Raphael, Daniel, raphaeld@bc.edu, Boston College
Reardon, Sean, sean.reardon@stanford.edu, Stanford University
Reckase, Mark, reckase@msu.edu, Psychometric Solutions
Reeves, Maggie, mreeves@urban.org, Urban Institute
Reichert-Schlax, Jasmin, jaschlax@uni-mainz.de, Johannes Gutenberg University Mainz
Ren, He, heren@uw.edu, University of Washington
Rho, Minjeong, minjeong019@gmail.com,
Rikoon, Samuel, srikoon@air.org, American Institutes for Research
Riley, Jack, jhebner2@huskers.unl.edu, University of Nebraska - Lincoln
Rios, Oscar, osrios@ucdavis.edu, University of California - Davis
Rodriguez, Michael, mcrdz@umn.edu, University of Minnesota
Rodriguez, Jennifer, jaracely@umich.edu, University of Michigan
Rome, Logan, lrome@cainc.com, Curriculum Associates
Roper, Donna, donna.roper@isd742.org,
Rose, Cedar, Cedar.Rose@mt.gov, Montana Office of Public Instruction
Rotou, Ourania, orotou@newmeridian.org, New Meridian Corporation
Roy, Sneha, sneharoy@ksu.edu, Kansas State University
Rozunick, Chris, christine.rozunick@tea.texas.gov, TEA
Runyon, Christopher, CRunyon@nbme.org, National Board of Medical Examiners
Rupp, Andre, arupp@nciea.org, Center for Assessment
Russell, Michael, michael.russell@bc.edu, Boston College
Sabboh, Godwin, gmsabboh@uncg.edu, University of North Carolina - Greensboro
Safitri, Shahnaz, ssafitri@purdue.edu,
Sahin, Fusun, fsahin@cainc.com, Curriculum Associates
Saiar, Amin, asaiar@aphanet.org, Board of Pharmacy Specialties
Sanchez, Edgar, suppression305@gmail.com,
Sandi, Erick, ericksandimonge@gmail.com,
Sarac, Merve, msarac@collegeboard.org, College Board

PARTICIPANT INDEX

(Last Name, First Name,Affiliation, Email)

Satkus, Paulius, satkus.paulius@gmail.com,
Sato, Edynn, edynn@satoeducationconsulting.com, Sato Education Consulting LLC
Sayin, Ayfer, ayfersayin@gazi.edu.tr, Gazi University
Scaruto, Matthew, Matt.Scaruto@atitesting.com, Ascend Learning
Schaefer, Katarina, schae2ke@jmu.edu, James Madison University
Schellman, Madeline, madeline.schellman@pearson.com, Pearson
Schewior, Lauritz, schewior@leibniz-ipn.de, IPN - Leibniz-Institute for Science and Mathematics Education
Schmidt, Katie, katie.a.schmidt@outlook.com, College Board
Schneider, Christina, cschne5934@aol.com, Cambium Assessment
Schroeder, Paul, paul.schroeder@copafs.org, Council of Professional Associations on Federal Statistics
Schwartz, Bob, schwartz.bob@gmail.com, National Conference of Bar Examiners
Schwartz, Danielle, dschwartz@cainc.com, Curriculum Associates
Sedhai, Khem, ksedhai@albany.edu, University at Albany, SUNY
Seo, Daeryong, daeryong.seo@pearson.com,
Sharairi, Sid, sid.sharairi@riversideinsights.com, Riverside Insights
Sharma, Deepak, phd21deepaks@iima.ac.in, IIM Ahemdabad
Shaw, Emily, eshaw@collegeboard.org, College Board
Shear, Benjamin, benjamin.shear@colorado.edu, University of Colorado Boulder
Shen, Qian, qian.shen1@ufl.edu, University of Florida
Shen, Ruiting, rs8422@nyu.edu, New York University
Shin, Seungwon, seungwon-shin-1@uiowa.edu, University of Iowa
Shin, Namsoo, namsoo@msu.edu,
Siebert, Julian, julian.siebert@ucsf.edu, University of California, San Francisco
Sinharay, Sandip, ssinharay@ets.org, ETS
Sireci, Stephen, sireci@acad.umass.edu, University of Massachusetts Amherst
Smith, Jessalyn, jsmith@datarecognitioncorp.com, DRC
Smith, Nate, nathaniel.ryan.smith@gmail.com, American Board of Internal Medicine
Snipes, Julie, jksnipes@udel.edu, University of Delaware
Soland, James, jgs8e@virginia.edu, University of Virginia
Solano-Flores, Guillermo, gsolanof@stanford.edu, Stanford University
Someshwar, Shonai, shonai.someshwar@gmail.com, National Council of State Boards of Nursing
Song, Hao, hsong@asppb.org, Association of State and Provincial Psychology Boards
Song, Dan, dan-song@uiowa.edu, The University of Iowa
Song, Yi, ysong@ets.org, ETS Research Institute
Spagnola, William, william.spagnola@nyu.edu, Curriculum Associates
Spangenberg, Bethany, Bethany.Spangenberg@azed.gov, Deputy Associate Superintendent of Assessment
Steedle, Jeffrey, jtsteedle@gmail.com,
Steiner, Peter, peter.steiner@phsg.ch, St.Gallen University of Teacher Education
Stoffers, Melissa, melissa.stoffers@unlv.edu, University of Nevada
Strambler, Michael, michael.strambler@yale.edu, Yale University
Student, Sanford, srstu@udel.edu, University of Delaware
Su, Yu-Lan, suyulan@gmail.com,
Su, Wei-Chia, weigasu@gmail.com, National Sun Yat-sen University
Suárez-Álvarez, Javier, suarezj@umass.edu, University of Massachusetts, Amherst
Suh, Hongwook, hongwooks@gmail.com, Cambium Assessment
suh, kyunghee, arkkhkimsuh@gmail.com, Cambium Assessment
Suk, Youmi, ysuk@tc.columbia.edu, Teachers College, Columbia University
Sun, Tianying, sty.9495@gmail.com,
Sussman, Joshua, jsussman@berkeley.edu, University of California, Berkeley
Swaby, Jeneve, swabyj@bc.edu, Boston College
Talreja, Vinita, vgtalrej@amazon.com,
Tan, Bin, btan4@ualberta.ca, University of Alberta
Tang, Xiuxiu, xtang8@nd.edu, University of Notre Dame
Tang, Nai-En, naientang@gmail.com, National Board of Chiropractic Examiners
Tang, Cheng, cheng.tang@uga.edu, The University of Georgia

PARTICIPANT INDEX

(Last Name, First Name,Affiliation, Email)

Tavares, Stephen, st4yk@virginia.edu, University of Virginia
Tenison, Caitlin, ctenison@ets.org, ETS
Tettey-Tawiah, Henrietta, hkt003@uark.edu,
Thacker, Arthur, athacker@humrro.org, Human Resources Research Organization
Thomas, Whitney, wthomas10@huskers.unl.edu, University of Nebraska - Lincoln
Thompson, Jake, wjakethompson@gmail.com, ATLAS, University of Kansas
Thompson Torbet, Jacquelyn, jackie.thompsontorbet@pearson.com,
Thurlow, Martha, marthathurlow001@gmail.com, Retried
Tian, Zewei, ztian27@uw.edu,
Tien, Ingrid, istien@g.ucla.edu, Centre for Addiction and Mental Health, McCain Centre for Child, Youth, and Family Mental Health
Tolentino, Lissette, ltolen@ufl.edu, University of Florida
Tolentino, Lissette, lissette.tolentino1@gmail.com, University of Central Florida
Tong, Ye, yetong@nbme.org, National Board of Medical Examiners
Toyama, Yukie, yukie.toyama@berkeley.edu,
Tran, Peter, petertran@umass.edu, University of Massachusetts, Amherst
Treadwell, Kelli, kelli.m.treadwell@gmail.com,
Trout, Nick, troutnic@msu.edu, Michigan State University
Tsai, Chia-Lin, chialin.tsai@unco.edu, University of Northern Colorado
Tsai, Tsung-hsun, ttsai@researchleague.org, Research League, LLC
Twing, Jon, jon.s.twing@icloud.com, HumRRO
Ulicheva, Anastasia, ana.ulicheva@pearson.com, Pearson
Valdivia Medinaceli, Montserrat, mvaldivia@cainc.com, Curriculum Associates
Valdivia Medinaceli, Montserrat, montse.bea.v.m@gmail.com, Curriculum Associates
Van Orman, Dustin, vanormd2@wwu.edu, Western Washington University
Veazey, Mary, mary.veazey@pearson.com,
Ventura, Claudia, claudia.j.ventura@uconn.edu, University of Connecticut
Vesey, Winona, winona.vesey@kaplan.com, Kaplan North America
Villafuerte, Catherina, catherina.villafuerte@uconn.edu, University of Connecticut
Vispoel, Walter, walter-vispoel@uiowa.edu, The University of Iowa
Vo, Yen, yen-vo@uiowa.edu, The University of Iowa
von Davier, Alina, avondavier@duolingo.com, Duolingo
Vu, Bui Nhat Anh, andy-vu@uiowa.edu, University of Iowa
Wackerle-Hollman, Alisha, wacke020@umn.edu, University of Minnesota
Walker, Michael, memwalker@gmail.com, HumRRO
Walker, Adrienne, adriennewalke5@hotmail.com, Ascend Learning
Walker, Michael, mwalker@humrro.org, Human Resources Research Organization (HumRRO)
Walsh, Cole, cwalsh@acuityinsights.com, Acuity Insights

Wan, Lei, lwan@collegeboard.org, College Board
Wang, Kuo, wangp@mail.smu.edu, Southern Methodist University
Wang, Ting, szpku.grady@gmail.com,
Wang, Zhuoran, wzhranran@gmail.com, NCSBN
Wang, Aijun, wajlm2003@gmail.com,
Wang, Bowen, bwang@nbce.org, National Board of Chiropractic Examiners
Wang, Yuan, ywang@ets.org, ETS
WANG, ZHEN, zwang@hkeaa.edu.hk, Hong Kong Examinations and Assessment Authority
Wang, Yu, yw3060@nyu.edu, New York University
Wang, Chun, wang4066@uw.edu, University of Washington
Wang, Huijuan, huijuan@email.sc.edu,
Wang, Ze, awsz@amazon.com, Amazon Web Services
Wang, Ting, TWang@theabfm.org, American Board of Family Medicine
Wang, Wenhao, wwh8623@gmail.com, HumRRO
WANG, XINYI, xinyiwangpsy@gmail.com,
Wang, Feng, hgwang98@gmail.com,
WANG, Teng, wangtengsh@126.com,
Wang, Zhaoyu, jessiewang@gatech.edu, Georgia Institute of Technology
Ward, Dylan, dward@abog.org, American Board of Obstetrics and Gynecology
Weeks, Jonathan, weeksjp@gmail.com, Stanford University
Wei, Hsin-Ro, esep23@gmail.com, Riverside Insights
Welch, Catherine, catherine-welch@uiowa.edu, University of Iowa
Wellberg, Sarah, sarah.wellberg@virginia.edu, University of Virginia
Welsh, Megan, welsh.megan@gmail.com,
Wesling, Piet, pwesling@albany.edu, University at Albany, SUNY
Wheeler, Jordan, jwheeler21@unl.edu, University of Nebraska - Lincoln
White, Lauren, law03k@gmail.com, Pearson
Whitfield, Erik, erwh9077@colorado.edu,
Whitmer, John, john@ld-insights.com, Learning Data Insights
Wild, Autumn, wildan@jmu.edu, James Madison University
Wiley, Andrew, awiley@acsventures.com, ACS Ventures, LLC
Williams, Kevin, kmwilliams@ets.org, ETS
Williams, Angie, amlong81@gmail.com, University of Kansas
Wind, Stefanie, swind@ua.edu, The University of Alabama
Wine, Marjorie, mwine122@gmail.com, Accessible Teaching, Learning and Assessment Systems (ATLAS), University of Kansas
Wireman, Jami, jwireman@newmeridian.org, New Meridian Corporation
Wise, Steven, stevewise23@gmail.com, EngagedMeasurement
Wolfe, Edward, ed@edwolfe.net, Iowa Testing Programs / University of Iowa
Wollack, James, jwollack@wisc.edu, University of Wisconsin - Madison
Workman, Trent, trent.workman@pearson.com, Pearson
Worsham, Hope, Hope.Worsham@ade.arkansas.gov, Arkansas Department of Education
Wry, Erin, erin.wry@bc.edu, Boston College
Wu, Tong, tong.wu@riversideinsights.com,
Wu, Yi-Fang, Yi-Fang.Wu@cambiumassessment.com, Cambium Assessment, Inc.
Wu, Lixin, lixinwu2@illinois.edu, University of Illinois at Urbana-Champaign
Wu, Sirui, sirui.wu@ubc.ca, University of British Columbia
Wu, Yi-Chen, wuxx0207@umn.edu, National Center on Educational Outcomes, University of Minnesota
Wu, Ruoqian, rw41@illinois.edu, University of Illinois Urbana-Champaign
Wyman, Austin, awyman@nd.edu, University of Notre Dame
Wyse, Adam, adam.wyse@renaissance.com, Renaissance
Xiao, Xingyao, xiaoxg@berkeley.edu, Stanford University
Xiao, Daihui, dxiao@msu.edu,
Xiong, Jiawei, jxiong@cainc.com, Curriculum Associates

PARTICIPANT INDEX

(Last Name, First Name,Affiliation, Email)

Xu, Yangmeng, yxu81@crimson.ua.edu, Pearson
Xu, Xiaochen, xixu@ucdavis.edu, University of California, Davis
Xu, Rujun, bwp7ab@virginia.edu, University of Virginia
Xu, Wei, x.wei1007@gmail.com, ISC2
Xue, Mingfeng, m_xue@uncg.edu, University of North Carolina Greensboro
Yan, Duanli, dyan08550@gmail.com,
Yaneva, Victoria, vyaneva@nbme.org, National Board of Medical Examiners
Yang, Ji Seung, jsyang@umd.edu, University of Maryland
Yang, Yi, yyang@ets.org, ETS
Yang, Yaxuan, yyaxuan@uga.edu, University of Georgia
Yang, Yiyao, yy3555@tc.columbia.edu, Teachers College, Columbia University
Yao, Yiting, yy21b@fsu.edu, Florida State University
Yasmin, Farzana, fyasmin@ualberta.ca, University of Alberta, Measurement, Evaluation & Data Science (MEDS)
Yavuz, Sinan, yavuzsinan@gmail.com, Amazon
Ye, Sangbeak, sye@fau.edu, Florida Atlantic University
Yin, Yunhang, yunhang@email.sc.edu, University of South Carolina
Young, Mackenzie, mackenzie.young@cambiumassessment.com,
Yu, Kefan, ky285@uw.edu,
Yuan, Ye, yyuan@gmac.com, GMAC
Zapata-Rivera, Diego, dzapata@ets.org, ETS
Zeng, Ji, zengj@michigan.gov,
Zenisky, April, azenisky@educ.umass.edu, University of Massachusetts, Amherst
Zhan, Peida, pdzhan@gmail.com, Zhejiang Normal University
Zhang, Yuxiao, zhan3971@purdue.edu, Purdue University
Zhang, Xiuyuan, xzhang@collegeboard.org, The College Board
Zhang, Yichi, yzhang97@usc.edu, Georgia Institute of Technology
Zhang, Yifan, zyf2020@connect.hku.hk, The University of Hong Kong
Zhang, Xiaowan, xiaowan@duolingo.com, Duolingo
zhang, ci, zhangci34@outlook.com,
Zhang, Jingru, jzhang2637@wisc.edu,
Zhang, Liru, lilyrelax88@gmail.com, Consultant
Zhao, Yu, tracy0227@gmail.com,
Zhao, Xinchu, xinchuz@gmail.com, Roblox
Zheng, Yi, yi.isabel.zheng@asu.edu, Arizona State University
Zhong, Xiaoting, xzhong6@uiowa.edu, University of Iowa
Zhou, Xinchang, xz77@illinois.edu,
Zhou, Xuechun, greenwhite0619@gmail.com, Ascend Learning
Zhou, Hao, haozhou@szu.edu.cn,
Zhu, Sizheng, sizhengzhu@outlook.com, Roblox
Zieher, Almut, almut.zieher@yale.edu, Education Collaboratory at Yale University
Zou, Tongtong, tz2345@tc.columbia.edu, Stanford University
Zwick, Rebecca, rzwick@cox.net, University of California, Santa Barbara

SPONSORS

PLATINUM SPONSORS

CollegeBoard

ets research institute

Pearson

GOLD SPONSORS

American Board of Internal Medicine®

Center for Assessment

SILVER SPONSORS

ACS VENTURES
BRIDGING THEORY & PRACTICE

ACT®

AAMC

Cambium Assessment

cognia™

Curriculum Associates
i-Ready

duolingo english test

edCount LLC®
because all students count

HARVARD
GRADUATE SCHOOL OF EDUCATION

HumRRO
HUMAN RESOURCES RESEARCH ORGANIZATION

NBME®

nwea

FRIEND SPONSORS

BUROS
CENTER FOR TESTING

UMass Amherst
College of Education

WebbAlign®
Powered by WCEPS

ADDITIONAL SPONSOR

VECTOR
PSYCHOMETRIC GROUP

LOS ANGELES

#NCME26

LinkedIn: ncme38

Facebook: NCMEPage

BlueSky: @ncme38

X/Twitter: @ncme38

National Council on Measurement in Education
520 S Walnut St. | Box 2388
Bloomington, IN 47402-2388
Phone: (812) 245-8096
Email: ncme@ncme.org