ncme

NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION

2023 ANNUAL MEETING

CHICAGO MARRIOTT DOWNTOWN

MAGNIFICENT MILE

April 12–15, 2023

Welcome from the Program Chairs

Welcome to the 2023 NCME Annual Meeting! We are so excited to have you join us virtually March 28-30 or in person April 12-15 in Chicago, Illinois.

The theme of the 2023 NCME Annual Meeting is Leveraging Measurement for Better Decisions. We do measurement to inform decisions. Decisions that consider data from good measurement practices are better decisions than those that do not. Many aspects of measurement are under attack right now. Some criticisms seem justified; some do not. We do need to improve our processes and tools. But we also need to be advocates for the appropriate use and application of the tools of our profession. We can make using measurement and assessment data 'cool' again. And we should. How can we do this?

We invite you to attend the training and presentation sessions and engage in conversations with other participants to find out how we can do this by making improvements in our processes and products, by communicating more effectively how data can be a force for good, by ensuring our use of data is a force for good, by being more collaborative both within and outside NCME, and by continuing to challenge, prod, encourage, question, and listen to each other. Whether you registered for the virtual or the in-person meeting, there are many wonderful sessions that you can attend. Below are just a small number of examples on a few selected topics from the wide variety of sessions.

Selected sessions on effective uses of test information and measurement models for better decisions:

- Using Eye Movement and Natural Language Processing to Inform Various Decisions (Thursday, March 30, 10:30am-12:00pm)
- Improving Assessment Decisions Using Collateral Information About Incorrect Responses and Response Times (Thursday, April 13, 8:00am-9:30am)
- Better Decisions Through Comprehensive Statistical Model Evaluation (Friday, April 14, 8:00am-9:30am)
- Combining Innovation and PAD to Economize Assessment Processes that Support Better Decisions (Saturday, April 15, 8:00am-9:30am)

Selected sessions that look back into the history, respond to current challenges, or look into the future of educational measurement:

- Historical Perspectives on Educational Measurement (Friday, April 14, 9:50am-11:20am)
- The Future is Now: Game-Changing Innovations in Educational Assessment (Friday, April 14, 8:00am-9:30am)
- Recent Evidence from the Pandemic and Test Optional Admissions Policies? (Thursday, April 13, 1:30pm-2:30pm)
- Challenges and Opportunities in Score Reporting (Thursday, March 30, 10:30am-12:00am)
- Challenges in Growth Measures and Accountability Decisions (Friday, April 14, 2:50pm-4:20pm)

Selected sessions on the impact of the COVID-19 pandemic:

- The Lingering Impact of the Pandemic from Multiple Analytic Perspectives (Thursday, March 30, 10:30am-12:00pm)
- Monitoring Performance of U.S. Students in the Pandemic with NAEP Long-Term Trend Assessments (Thursday, April 13, 11:40am-1:10pm)
- The Impact of Pandemic on Testing Industry (Saturday, April 15, 4:40pm-6:10pm)

Selected sessions on culturally relevant and culturally responsive assessment:

- Developing Culturally Relevant Assessment Content: Lessons Learned and the Road Ahead (Thursday, March 30, 10:30am-12:00pm)
- Culturally Responsive and Related Approaches to Assessment: What are They? (Thursday, April 13, 11:40am-1:10pm)

Selected sessions grappling with equity and inclusion in the measurement profession and in the larger society:

- State of the Field: Gender and Racial Equity in Educational Measurement (Saturday, April 15, 9:50am to 11:20am)
- Improving Measures of Opportunity to Learn (OTL) to Address Systemic Inequity (Saturday, April 15, 2:50pm-4:20pm)
- Test Equity and Fairness from the Voices that Matter (Saturday, April 15, 4:40pm-6:10pm)

Selected sessions showcasing technical advances in various areas of the field:

- Research Blitz: Advances in Item Response Theory (Wednesday, March 29, 2:45pm-3:45pm)
- Latest Work in Item Difficulty Modeling and Cognitive Complexity (Thursday, April 13, 9:50am-11:20am)
- Using New Techniques to Gather Validity Evidence (Friday, April 14, 9:50am-11:20am)
- Cheating Detection Using Machine Learning and Deep Learning Methods (Saturday, April 15, 9:50am-11:20am)

We'd also like to call your attention to the electronic board (eBoard) sessions and a few special sessions. The eBoard sessions run all day on Thursday April 13, including the Graduate Student Issues Committee (GSIC) eBoard sessions and the clustered eBoard sessions. The clustered eBoard is a new format that we started this year where two or three Presenters with similar topics share the same presentation station so that there can be more interactions among the Presenters. The 2023 edition of the NCME Gala Comedy Event takes place 4:40pm-6:10pm on Thursday April 13. In the 2023 NCME Career Award Session (Friday, April 14, 1:30pm-2:30pm) Dr. William Stout will address his major accomplishments in formative assessments. Last but not least, one session is devoted to remembering and celebrating the contributions of Ronald K. Hambleton, one of the most influential and productive psychometricians – Remembering Ron: Reflections on a Career and a Legacy (Saturday, April 15,
4:40pm-6:10pm).

We are so thankful to all who have contributed to this engaging program, including authors of the submissions and all volunteers. We are appreciative of the reviewers for providing helpful feedback as well as colleagues who volunteered as discussants and chairs. We want to thank Qing Yi, Nathan Wall, and Alfonso Martinez, the Training and Professional Development Committee Chairs, as well as Sergio Araneda and Janine A. Jackson, chairs of the Graduate Student Issues Committee and Can Shao, chair of the Committee on Diversity Issues in Testing for their work on the program. We are also very thankful to the previous NCME program chairs for their help. Finally, we thank the NCME President Deborah Harris for her time, patience, encouragement, and continuous support.

Dongmei Li, Wei Tao, and Alexis Oakley
2023 NCME Annual Meeting Co-Chairs

Table of Contents

General Meeting Information 06
Floor Plans 07
NCME Board of Directors 09
Proposal Reviewers 11
Schedule at a Glance 13

Virtual Sessions

Tuesday, March 28 24
Wednesday, March 29 26
Thursday, March 30 36

In-Person Sessions

Wednesday, April 12 46
Thursday, April 13 49
Friday, April 14 93
Saturday, April 15 113

General Meeting Information

Welcome to the 2023 NCME Annual Meeting in Chicago, IL!

NCME REGISTRATION & INFORMATION DESK

The NCME Registration & Information Desk is located on the 5th floor in the Registration Booth area at the Chicago Marriott Downtown Magnificent Mile. Stop by the registration desk to pick up your conference materials including your name badge. Stop by the information desk to ask questions about your membership, the program, the NCME Events app, and for any other questions! If you are participating in the NCME 5K Fun Run, please stop by the desk to pick up your shirt.

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

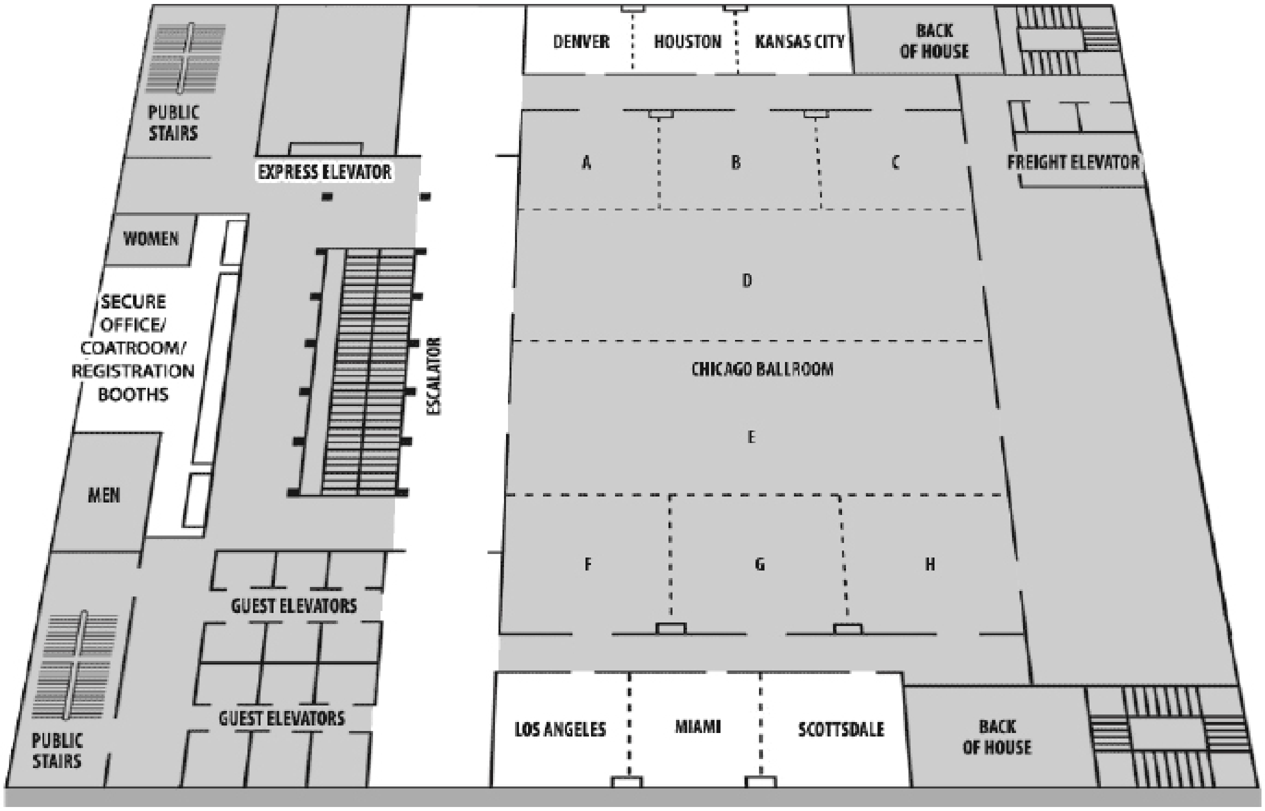
Wednesday, April 12	**7:30 am – 5:00 pm**
Thursday, April 13	**7:30 am – 5:00 pm**
Friday, April 14	**7:30 am – 5:00 pm**
Saturday, April 15	**7:30 am – 12:00 pm**

TWITTER

Share your experience at the NCME Annual Meeting by using #NCME2023

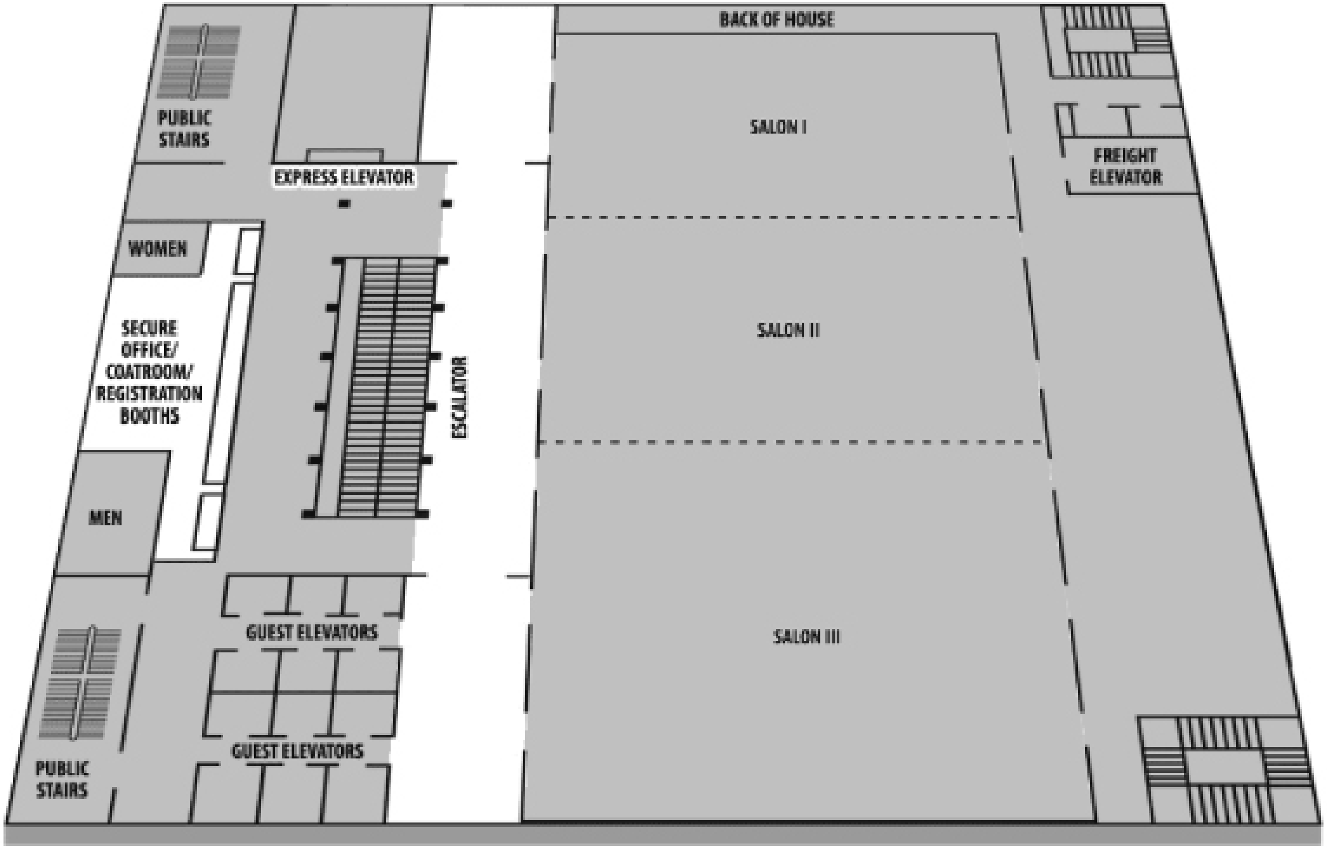
Floor Plans

5th Floor - Chicago Marriott Downtown Magnificent Mile



Floor Plans

7th Floor - Chicago Marriott Downtown Magnificent Mile



NCME Board of Directors

Deborah J. Harris (President)
University of Iowa

Michael E. Walker (President Elect)
ETS

Derek Briggs (Past President)
University of Colorado Boulder

Antoinette Stroter
Chesterfield County Public Schools

Sharyn Rosenberg
National Assessment Governing Board

Li Cai
UCLA-CRESST

Susan Davis-Becker
ACS Ventures

Kyndra Middleton
Howard University

Ellen Forte
edCount

Editors

Journal of Educational Measurement	**Chun Wang** *University of Washington*
Educational Measurement: Issues and Practice	**Zhongmin Cui** *CFA Institute*
ITEMS Editor	**Brian C. Leventhal** *James Madison University*
NCME Book Series Editor	**Kadriye Ercikan** *Educational Testing Service*
NCME Newsleter Editor	**Arthur Thacker** *HumRRO*
NCME Website Editors	**Erin Banjanovic** *Curriculum Associates* **Jing Miao** *Educational Testing Service*

2023 Annual Meeting Chairs

Annual Meeting Program Chairs	Dongmei Li *ACT, Inc.*
	Wei Tao *Cambium Assessment, Inc.*
	Alexis Oakley *University of Iowa*
Training and Professional Development Committee Chairs	Qing Yi *Pearson*
	Nathan Wall *eMetric*
	Alfonso Martinez *University of Iowa*
Fitness Run/Walk Directors	Jill R. van den Heuvel *Consultant*
	Katherine Furgol Castellano *Educational Testing Service*
	Brian F. French *Washington State University*

Proposal Reviewers

Terry Ackerman
Tony Albano
Jeff M. Allen
Benjamin Andrews
Serkan Arikan
Nana Amma Asamoah
Nafisa Awwal
Elizabeth Ayers-Wright
Mariana Barragan Torres
Deni Basaraba
Bozhidar M. Bashkov
Michael Beck
Douglas F. Becker
Stefan Behrendt
Yoav Bergner
Katrina Borowiec
Donna J Butterbaugh
Luciana Cancado
Maritza Casas
Roti Chakraborty
Hong Chen
Jianshen Chen
Troy Chen
Yiling Cheng
Yi-Chen Chiang
Edison M. Choe
Hye-Jeong Choi
Amy Clark
Whitney Smiley Coggeshall
Kimberly Colvin
Jeffrey Cucina
Zhongmin Cui
Laurie Davis
Susan Davis-Becker
Teresa Dawber
Christine DeMars
Onur Demirkaya
Jiayi Deng
Khagendra Raj Dhakal
Ismail Dilek
Kerry Englert
Fen Fan
Meng Fan
Rich Feinberg
Steve Ferrara
Leah Feuerstahler
Anthony D. Fina
Holmes Finch
Yanyan Fu
Robert Thomas Furter
Yizhu Gao
Terri Gibbs-Burke
Sakine Gocer Sahin

Brian Gong
Joshua Goodman
Kylie Gorney
Raman Grover
Yage Guo
Brian Habing
Sarah Hagge
Peter Halpin
Polina Harik
Samuel Haring
Deborah J Harris
Qiwei He
Wei He
Yong He
Ian Hembry
Amy Hendrickson
Fiona Hinds
Michelle Hock
Kari Hodge
Thomas P. Hogan
Hyeri Hong
Minju Hong
Anne Corinne Huggins-Manley
Janine Jackson
Xuejun Ryan Ji
Yue Jia
Kuan Yu Jin
Mark Johnson
Edmund Jones
Radhika Kapoor
Yusuf Kara
Hacer Karamese
Justin L. Kern
Kyung Yong Kim
Seongeun Kim
Stella Kim
Sunhee Kim
Young Yee Kim
Yun-Kyung Kim
Jennifer Kobrin
Audra Kosh
Kevin Krost
Olga Kunina-Habenicht
Alexander Kwako
Hollis Lai
Meredith Langi
Haeju Lee
Yongseok Lee
Brian C Leventhal
Dongmei Li
Isaac Li
Xueming Li
Min Liang

Yuan-Ling Liaw
Ye Lin
Chunyan Liu
Huan Liu
Liu Liu
Susan Lottridge
Quentin Ulysses Adrian Love
Ru Lu
Jinwen Luo
Yong Luo
Weicong Lyu
Hotaka Maeda
Jaime Malatesta
Yue Mao
Xia Mao
Scott Marion
Jose Felipe Martinez
Njideka Gertrude Mbelede
Martha McCall
Janet Mee
Alejandra Miranda
Scott Monroe
Kristin M. Morrison
Megan M Mulvihill
Aaron Myers
Xinyu Ni
Kyle Nickodem
Katherine Nolan
Francis O'Donnell
Maura O'Riordan
Yesim Ozer Ozkan
Yiqin Pan
Seohee Park
Thanos Patelis
Chris Patterson
Richard Patz
Michael R Peabody
Duy N. Pham
Cornelis Potgieter
Sonya Powers
Stanley N Rabinowitz
Ray E. Reichenberg
Kelly Rewley
Michael C. Rodriguez
Jonathan Rubright
Leslie Rutkowski
Alper Sahin
Edgar Sanchez
Merve Sarac
Shimon Sarraf
Paulius Satkus
Edynn Sato
Amy Schmidt

Christina Schneider
Matthew Schultz
Robert Schwartz
Yelisey A Shapovalov
Benjamin R. Shear
Yawei Shen
Mark David Shermis
Philipp Sonnleitner
Dorota Staniewska
Jeffrey Steedle
Elizabeth Stone
Sanford Student
Yu-Lan Su
Joshua Sussman
Kimberly Swygert
Shuqin Tao
Wei Tao
Brad Thiessen
Kathryn Nicole Thompson
Lissette Tolentino
Ye Tong
Anna Topczewski
Emily Karen Toutkoushian
Mubeshera Tufail
Stephanie Underhill
Nazli Uygun
Montserrat B Valdivia Medinaceli
Nathan Wall
Gabriel Wallin
Bowen Wang
Zachary Warner
Jonathan Weeks
Alexander Weissman
Sarah Wellberg
Andrew Wiley
Phoebe C Winter
Tarid Wongvorachan
Yi-Fang Wu
Yi-jung Wu
Adam E Wyse
Jiawei Xiong
Guanlan Xu
Menglin Xu
Lihua Yao
Guler Yavuz
Qing Yi
Hanwook Yoo
Paul Zavitkovsky
Mingqin Zhang
Xiuyuan Zhang
Mingying Zheng
Xiaoting Zhong
Xuechun Zhou
Anna Zilberberg
Selay Zor

Training Session Reviewers

Alfonso Martinez
Yale Quan
Nathan Wall
Chun Wang
Jonathan Weeks
Qing Yi

Graduate Student Abstract Reviewers

Serkan Arikan
Nafisa Awwal
Elizabeth Ayers-Wright
Deni Basaraba
Michael Beck
Magdalen Beiting-Parrish
Katrina Borowiec
Laurie Davis
Jiayi Deng
Khagendra Raj Dhakal
Rich Feinberg
Yanyan Fu
Sakine Gocer Sahin
Guher Gorgun
Samuel Haring
Hyeri Hong
Janine Jackson
Xuejun Ryan Ji
Radhika Kapoor
Stella Kim
Yun-Kyung Kim
Kevin Krost
Olga Kunina-Habenicht
Alexander Kwako
Min Liang
Huan Liu
Liu Liu
Jinwen Luo
Weicong Lyu
Yue Mao
Kristin M. Morrison
Maura O'Riordan
Yesim Ozer Ozkan
Cornelis Potgieter
Onur Ramazan
Edgar Sanchez
Merve Sarac
Shimon Sarraf
Christina Schneider
Yelisey A Shapovalov
Sanford Student
Wei Tao
Nazli Uygun
Montserrat B Valdivia Medinaceli
Bowen Wang
Tarid Wongvorachan
Mingying Zheng
Xuechun Zhou
Selay Zor

Schedule at a Glance

(Sessions marked with * are training sessions with additional cost.)

Tuesday, March 28

Begin Time	End Time	Room	Session Title
8:45 am CT	12:45 pm CT	Virtual	Diagnostic Classification Models: Advanced Applications*
8:45 am CT	12:45 pm CT	Virtual	Using Stan for Bayesian Psychometric Modeling (Part I)*
1:00 pm CT	5:00 pm CT	Virtual	Optimal Test Design Approach to Fixed and Adaptive Test Construction using R*
1:00 pm CT	5:00 pm CT	Virtual	Tools and Strategies for the Design and Evaluation of Interactive Dashboard Reports*
1:00 pm CT	5:00 pm CT	Virtual	Tools For Analyzing NAEP and TIMSS Data in R Using Latent Regression*

Wednesday, March 29

Begin Time	End Time	Room	Session Title
8:45 am CT	12:45 pm CT	Virtual	An Overview of Operational Psychometric Work in Real World*
8:45 am CT	12:45 pm CT	Virtual	Applying Data Mining Methods to Detect Test Fraud*
8:45 am CT	12:45 pm CT	Virtual	Using Stan for Bayesian Psychometric Modeling (Part II)*
1:00 pm CT	2:30 pm CT	Virtual	Comparison and Integration of Generalizability Theory with Structure Equating Modeling
1:00 pm CT	2:30 pm CT	Virtual	Automated Assessment of Writing and Reading Proficiency
1:00 pm CT	2:30 pm CT	Virtual	Models and Applications of Process Data and Eye Movement
1:00 pm CT	2:30 pm CT	Virtual	The Impact of COVID-19 and Learning Recovery
1:00 pm CT	2:30 pm CT	Virtual	Approaches for Evaluating and Reporting Strength of Validity Evidence for Assessments
2:45 pm CT	3:45 pm CT	Virtual	Platforms and Strategies to Enhance Learning
2:45 pm CT	3:45 pm CT	Virtual	Research Blitz: Advances in Item Response Theory
2:45 pm CT	3:45 pm CT	Virtual	Standard Setting and Proficiency Level Descriptors
2:45 pm CT	3:45 pm CT	Virtual	GSIC Virtual eBoard Session
4:00 pm CT	5:00 pm CT	Virtual	Factor Analysis Model Fit
4:00 pm CT	5:00 pm CT	Virtual	Demonstrations: Software and Training Module

Wednesday, March 29

Begin Time	End Time	Room	Session Title
4:00 pm CT	5:00 pm CT	Virtual	The Roles of Distractors in Developing Digital Assessments Within Assessment Engineering Frameworks
4:00 pm CT	5:00 pm CT	Virtual	Virtual eBoard Session
5:15 pm CT	6:45 pm CT	Virtual	Presenting from Three Continents on Three Topics: Collaborative Problem Solving, Linked Scores, and Propensity Score Estimation

Thursday, March 30

Begin Time	End Time	Room	Session Title
8:45 am CT	10:15 am CT	Virtual	Identification of Low-Effort Responses and Measurement of Digital Literacy
8:45 am CT	10:15 am CT	Virtual	Multidimensionality and Adaptive Learning Modeling
8:45 am CT	10:15 am CT	Virtual	Test Equating and Linking Challenges and New Methodology
8:45 am CT	10:15 am CT	Virtual	Cognitive Diagnosis Models and Practice
8:45 am CT	10:15 am CT	Virtual	Methods and Applications of Survey Research and Noncognitive Assessment
10:30 am CT	12:00 pm CT	Virtual	Using Eye Movement and Natural Language Processing to Inform Various Decisions
10:30 am CT	12:00 pm CT	Virtual	Developing Culturally Relevant Assessment Content: Lessons Learned and the Road Ahead
10:30 am CT	12:00 pm CT	Virtual	Quality Implications of Assessment Engineering in Developing Digital Applications of Testing and Learning
10:30 am CT	12:00 pm CT	Virtual	The Lingering Impact of the Pandemic from Multiple Analytic Perspectives
10:30 am CT	12:00 pm CT	Virtual	[NCME Book Series] Challenges and Opportunities in Score Reporting
1:00 pm CT	2:30 pm CT	Virtual	Investigations to Inform Item Pools and Test Design
1:00 pm CT	2:30 pm CT	Virtual	Differential Item Functioning: Sources and Detection
1:00 pm CT	2:30 pm CT	Virtual	Beyond Basketball and Bodegas: Pursuing True Cultural Validity in Formative Assessment
1:00 pm CT	2:30 pm CT	Virtual	Linking and Equating: Models and Tradeoffs between Sample and Precision
3:00 pm CT	4:30 pm CT	Virtual	Leverage the Partially Confirmatory Approach to Psychometric Modeling with Bayesian Regularization
3:00 pm CT	4:30 pm CT	Virtual	Foundational Competencies in Educational Measurement: How Do Measurement Careers Require Foundational Competencies?

Wednesday, April 12

Begin Time	End Time	Room	Session Title
8:00 am CT	12:00 pm CT	Chicago Ballroom A	An Introduction to Bayesian Statistics*
8:00 am CT	12:00 pm CT	Chicago Ballroom G	Demystify Amazon Web Services (AWS): Cloud Computing, and Psychometric Applications*
8:00 am CT	12:00 pm CT	Chicago Ballroom H	Professional Training for Graduate Students in Measurement*
8:00 am CT	5:00 pm CT	Chicago Ballroom B	Addressing the Data Challenges of Next-generation Assessments: Data Science Upskilling for Psychometricians*
8:00 am CT	5:00 pm CT	Chicago Ballroom F	Bayesian Networks in Educational Assessment (Book by Springer)*
1:00 pm CT	5:00 pm CT	Chicago Ballroom A	Embedded Standard Setting in Practice*
1:00 pm CT	5:00 pm CT	Chicago Ballroom C	An Introduction to Creating Video Games for Measurement: From Design to Analysis*
1:00 pm CT	5:00 pm CT	Chicago Ballroom G	Visualizations and Interactive Graphics using R*
1:00 pm CT	5:00 pm CT	Chicago Ballroom H	Sequence Mining Methods on Process Data in Large-Scale Assessments*
4:00 pm CT	7:00 pm CT	Old Town	Board Meeting

Thursday, April 13

Begin Time	End Time	Room	Session Title
8:00 am CT	9:30 am CT	Chicago Ballroom B/C	Implementing More Student-Centric Measurement Processes: Adventures in Developing the Digital SAT
8:00 am CT	9:30 am CT	Chicago Ballroom D [Recorded]	Empowering Process Data for Data-Informed Decision-Making in Measurement
8:00 am CT	9:30 am CT	Chicago Ballroom E [Recorded]	Improving Assessment Decisions Using Collateral Information About Incorrect Responses and Response Times
8:00 am CT	9:30 am CT	Chicago Ballroom F	Assessing Collaborative Problem Solving at Scale: Individual Contribution to Teamwork
8:00 am CT	9:30 am CT	Chicago Ballroom G/H	Research Blitz: IRT Models
8:00 am CT	9:30 am CT	Denver/Houston	Moving Towards an Equitable and Just Profession: Lessons Learned from The Field
8:00 am CT	9:30 am CT	Los Angeles/ Miami	Design and Evaluation of Adaptive Testing in Large-Scale Survey Assessments
8:00 am CT	9:30 am CT	Salon I	eBoard Session 1
9:50 am CT	11:20 am CT	Chicago Ballroom B/C	Advances in Item Response and Response Time Modeling

Thursday, April 13

Begin Time	End Time	Room	Session Title
9:50 am CT	11:20 am CT	Chicago Ballroom D [Recorded]	Issues and Strategies in Maintaining Testing Programs Internationally and in Various Languages
9:50 am CT	11:20 am CT	Chicago Ballroom E [Recorded]	Latest Work in Item Difficulty Modeling and Cognitive Complexity
9:50 am CT	11:20 am CT	Chicago Ballroom F	Cognitive Diagnostic Modeling: Mathematical Issues and Model Specifications
9:50 am CT	11:20 am CT	Chicago Ballroom G/H	Internships in the Measurement Profession: A Discussion Among Organizers, Mentors, and Students
9:50 am CT	11:20 am CT	Denver/ Houston	Automatic Generated Items and Automatic Enemy Item Detection
9:50 am CT	11:20 am CT	Salon I	Clustered eBoard Session 1
11:40 am CT	1:10 pm CT	Chicago Ballroom B/C	Supporting Test Security of Remote Testing with Process Data Analytics and AI
11:40 am CT	1:10 pm CT	Chicago Ballroom D [Recorded]	Measuring Change in a Changing World: Updating Frameworks without Breaking Trends
11:40 am CT	1:10 pm CT	Chicago Ballroom E [Recorded]	Monitoring Performance of U.S. Students in the Pandemic with NAEP Long-Term Trend Assessments
11:40 am CT	1:10 pm CT	Chicago Ballroom F	Measurement Models for the Purpose of Evaluating Interventions and Programs
11:40 am CT	1:10 pm CT	Chicago Ballroom G/H	Meeting the Challenge: The Law School Admission Test in Changing Times
11:40 am CT	1:10 pm CT	Denver/ Houston	Culturally Responsive and Related Approaches to Assessment: What are They?
11:40 am CT	1:10 pm CT	Los Angeles/ Miami	Assessing the Impact of Feedback in Computer-Based Assessments
11:40 am CT	1:10 pm CT	Salon I	GSIC eBoard Session 1
1:30 pm CT	2:30 pm CT	Chicago Ballroom B/C	Culturally Responsive Assessment: Method and Impact
1:30 pm CT	2:30 pm CT	Chicago Ballroom D [Recorded]	Recent Evidence from the Pandemic and Test Optional Admissions Policies?
1:30 pm CT	2:30 pm CT	Chicago Ballroom E [Recorded]	Content-Referenced Growth: Creating Instructionally Actionable Growth Interpretations in Reading and Mathematics Assessments
1:30 pm CT	2:30 pm CT	Chicago Ballroom F	Differential Item Functioning Detection Methods
1:30 pm CT	2:30 pm CT	Chicago Ballroom G/H	Predicting Item Difficulty of Language Tests

Thursday, April 13

Begin Time	End Time	Room	Session Title
1:30 pm CT	2:30 pm CT	Denver/ Houston	Re-thinking Construct Definitions and Measurement Methods to Include Black and Hispanic Cultures
1:30 pm CT	2:30 pm CT	Los Angeles/ Miami	Research Blitz: Various Uses of Process Data
1:30 pm CT	2:30 pm CT	Salon I	GSIC eBoard Session 2
2:50 pm CT	4:20 pm CT	Chicago Ballroom B/C	Innovating Assessments: Towards Next Generation Assessments of 21st Century Skills
2:50 pm CT	4:20 pm CT	Chicago Ballroom D [Recorded]	Practical Applications of NLP and Text Mining Techniques for Test Development Tasks
2:50 pm CT	4:20 pm CT	Chicago Ballroom E [Recorded]	Psychometric Implications of Item Exposure in Standardized Testing: Investigative Procedures and Impact
2:50 pm CT	4:20 pm CT	Chicago Ballroom F	Research Blitz: Automated Scoring
2:50 pm CT	4:20 pm CT	Chicago Ballroom G/H	Research Blitz: Impact of COVID-19
2:50 pm CT	4:20 pm CT	Denver/ Houston	Differential Item Functioning Detection and More
2:50 pm CT	4:20 pm CT	Los Angeles/ Miami	Issues in the Use of Anonymous Population Data to Infer Learning from Gameplay
2:50 pm CT	4:20 pm CT	Salon I	Clustered eBoard Session 2
4:40 pm CT	6:10 pm CT	Chicago Ballroom B/C	Through-Year Assessment and Growth
4:40 pm CT	6:10 pm CT	Chicago Ballroom D [Recorded]	Gala NCME Comedy Event - 2023
4:40 pm CT	6:10 pm CT	Chicago Ballroom E [Recorded]	[SIGIMIE Session] Diagnostic Measurement: Operational and Implementational Issues
4:40 pm CT	6:10 pm CT	Chicago Ballroom F	Assessing Non-cognitive Traits with Multi-dimensional Forced-choice Assessments: Design, Development, and Validation
4:40 pm CT	6:10 pm CT	Chicago Ballroom G/H	Computer Adaptive Testing: Variations and Impacts
4:40 pm CT	6:10 pm CT	Denver/ Houston	GSIC Standards Study Group: Recommendations from Graduate Students for Its New Version
4:40 pm CT	6:10 pm CT	Los Angeles/ Miami	Methodological Advances in Detecting and Accounting for Noneffortful Responding
4:40 pm CT	6:10 pm CT	Salon I	eBoard Session 2

Friday, April 14

Begin Time	End Time	Room	Session Title
6:00 am CT	7:00 am CT	Meet in Hotel Lobby	NCME Fitness Run/Walk
8:00 am CT	9:30 am CT	Chicago Ballroom B/C	Computer Adaptive Testing: Item Pool Development and Calibration
8:00 am CT	9:30 am CT	Chicago Ballroom D [Recorded]	Better Decisions Through Comprehensive Statistical Model Evaluation
8:00 am CT	9:30 am CT	Chicago Ballroom E [Recorded]	The Future is Now: Game-Changing Innovations in Educational Assessment
8:00 am CT	9:30 am CT	Chicago Ballroom F	Integrating Process Data in Psychometric Models
8:00 am CT	9:30 am CT	Chicago Ballroom G/H	Advanced Technology Use in TIMSS and PIRLS
8:00 am CT	9:30 am CT	Denver/ Houston	Method and Conceptual Development in Test Scaling, Linking, and Equating
8:00 am CT	9:30 am CT	Los Angeles/ Miami	Classroom and Instructionally Embedded Assessment
9:50 am CT	11:20 am CT	Chicago Ballroom B/C	Challenges in Online Testing and/or Online Proctoring
9:50 am CT	11:20 am CT	Chicago Ballroom D [Recorded]	Historical Perspectives on Educational Measurement
9:50 am CT	11:20 am CT	Chicago Ballroom E [Recorded]	Using New Techniques to Gather Validity Evidence
9:50 am CT	11:20 am CT	Chicago Ballroom F	Test Security
9:50 am CT	11:20 am CT	Chicago Ballroom G/H	Using Measurement to Improve Educational Decisions
9:50 am CT	11:20 am CT	Denver/ Houston	Advances in Item Response Modeling
9:50 am CT	11:20 am CT	Los Angeles/ Miami	Standard Setting
11:40 am CT	1:10 pm CT	Chicago Ballroom B/C	Predicting Item Difficulty and Response Latencies
11:40 am CT	1:10 pm CT	Chicago Ballroom D [Recorded]	[SIGIMIE Session] Leveraging Process Data to Better Understand Engagement and Motivation in Large-Scale Assessment
11:40 am CT	1:10 pm CT	Chicago Ballroom E [Recorded]	Putting Humpty Dumpty Back Together: Practical Advice for Synthesizing Validity Evidence
11:40 am CT	1:10 pm CT	Chicago Ballroom F	Cognitive Diagnostic Assessment: Modeling and Design

Friday, April 14

Begin Time	End Time	Room	Session Title
11:40 am CT	1:10 pm CT	Chicago Ballroom G/H	Development and Methodologies for Operational CAT Programs with Advanced Requirements
11:40 am CT	1:10 pm CT	Denver/ Houston	Causal Modeling of Log Data from EdTech
11:40 am CT	1:10 pm CT	Los Angeles/ Miami	Advancing Psychometric Processes and Tools in a Changing Testing Environment
1:30 pm CT	2:30 pm CT	Chicago Ballroom B/C	Comparability of Scores from Through-Year and Traditional State Assessments: Examining Louisiana
1:30 pm CT	2:30 pm CT	Chicago Ballroom D [Recorded]	Impact of College Admission Test Mandate and Alternative Approaches
1:30 pm CT	2:30 pm CT	Chicago Ballroom E [Recorded]	2023 NCME Career Award Session William Stout, From Martingales to Formative Assessments (FAs): A Career in Progress
1:30 pm CT	2:30 pm CT	Chicago Ballroom F	Demonstrations: Session 1
1:30 pm CT	2:30 pm CT	Chicago Ballroom G/H	Innovations in Assessment and Feedback
1:30 pm CT	2:30 pm CT	Denver/Houston	The Development and Utility of Learning Progressions in the K-12 Setting
1:30 pm CT	2:30 pm CT	Los Angeles/ Miami	Establishing Instructionally Meaningful Cut Scores with Embedded Standard Setting
1:30 pm CT	2:30 pm CT	Salon I	Reception for Researchers from Historically Marginalized Groups
2:50 pm CT	4:20 pm CT	Chicago Ballroom B/C	Challenges in Growth Measures and Accountability Decisions
2:50 pm CT	4:20 pm CT	Chicago Ballroom F	Rater Effect Evaluation and Mitigation
2:50 pm CT	4:20 pm CT	Chicago Ballroom G/H	Expanding the Conceptualization of Fairness for Digital Learning and Assessment
2:50 pm CT	4:20 pm CT	Denver/ Houston	Advances in Language Assessment
2:50 pm CT	4:20 pm CT	Los Angeles/ Miami	Data-driven Analysis of Latent Structures for Cognitive Diagnosis Models in Educational Assessments
4:40 pm CT	6:15 pm CT	Chicago Ballroom D/E [Recorded]	Business Meeting and Presidential Address
6:30 pm CT	8:30 pm CT	Salon I and II	President’s Reception (all NCME attendees are welcome)

Saturday, April 15

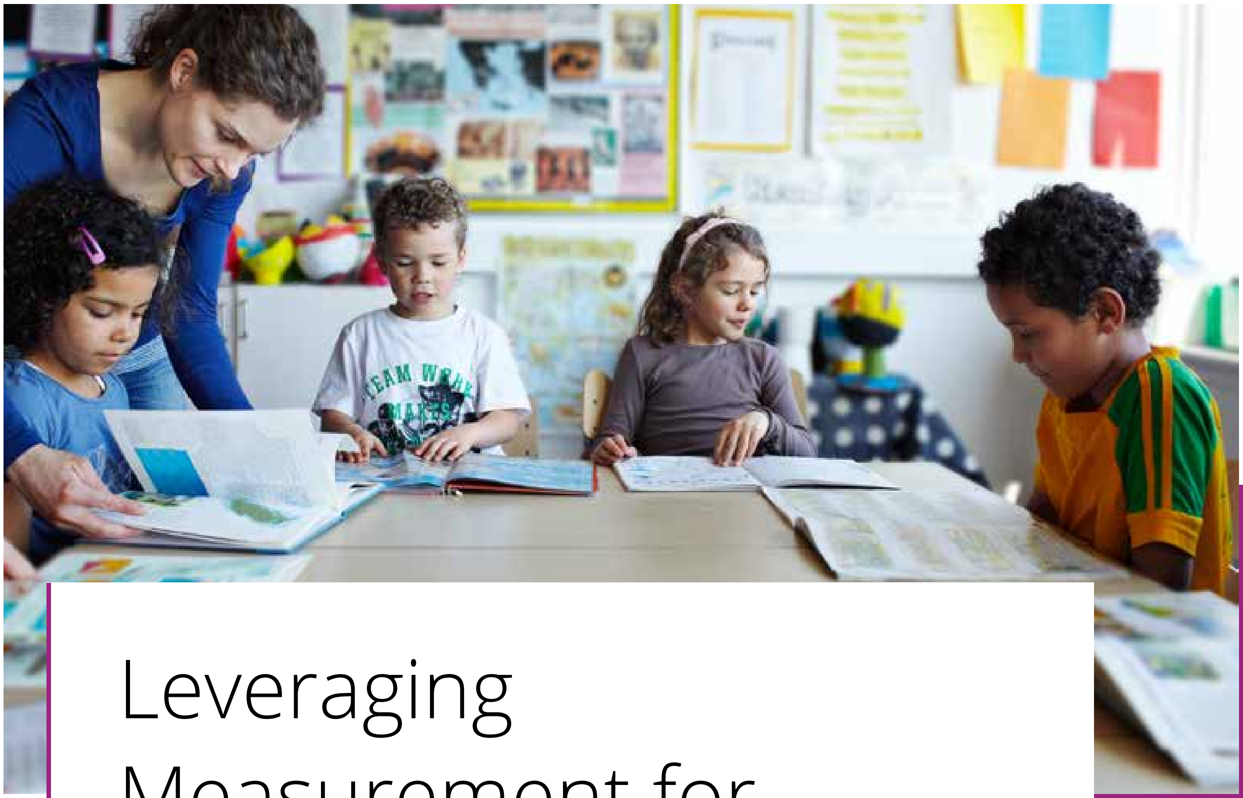
Begin Time	End Time	Room	Session Title
8:00 am CT	9:30 am CT	[AERA HOTEL] Inter-Continental Chicago Magnificent Mile: Floor 4th - Camelot Room	[Joint Session with AERA Division D] Examining AI and Machine Learning Through a Fairness and Equity Lens
8:00 am CT	9:30 am CT	Chicago Ballroom B/C	Combining Innovation and PAD to Economize Assessment Processes that Support Better Decisions
8:00 am CT	9:30 am CT	Chicago Ballroom D [Recorded]	Automated Test Assembly in Operational Assessment Programs
8:00 am CT	9:30 am CT	Chicago Ballroom E [Recorded]	Foundational Competencies in Educational Measurement: NCME Task Force Consensus and Debate
8:00 am CT	9:30 am CT	Chicago Ballroom F	Innovative Methodologies in Computational Statistics
8:00 am CT	9:30 am CT	Denver/ Houston	Use of Metrics and Thresholds in AI Scoring Model Evaluation
8:00 am CT	9:30 am CT	Los Angeles/Miami	[SIGIMIE Session] Advancing Perspectives on Practice Analysis for Credentialing Examinations
9:50 am CT	11:20 am CT	Chicago Ballroom B/C	[Joint Session with AERA Division D] State of the Field: Gender and Racial Equity in Educational Measurement
9:50 am CT	11:20 am CT	Chicago Ballroom D [Recorded]	Cheating Detection Using Machine Learning and Deep Learning Methods
9:50 am CT	11:20 am CT	Chicago Ballroom E [Recorded]	Holistic Admissions With Test-Optional Policies: Application Essays, Recommendation Letters, and Other Factors
9:50 am CT	11:20 am CT	Chicago Ballroom F	Analytics and Design Considerations to Inform Test Development
9:50 am CT	11:20 am CT	Chicago Ballroom G/H	The Digital SAT: the Impact of Changes
9:50 am CT	11:20 am CT	Denver/ Houston	Research Blitz: On Various Topics from Test Design and Scale Validation to Modeling of Response Bias and Missing Data
9:50 am CT	11:20 am CT	Los Angeles/Miami	Test Security Breaches: Prevalence, Detection Strategies, and Decision Making
11:40 am CT	1:10 pm CT	Chicago Ballroom B/C	Test Comparability around the World: Methodological Challenges and Solutions
11:40 am CT	1:10 pm CT	Chicago Ballroom D [Recorded]	New Approaches to Some Contemporary Problems in Evaluating Achievement and Growth
11:40 am CT	1:10 pm CT	Chicago Ballroom E [Recorded]	[SIGIMIE Session] Big Data, Big Change, Big Decision

Saturday, April 15

Begin Time	End Time	Room	Session Title
11:40 am CT	1:10 pm CT	Chicago Ballroom F	Validating a Writing Trait Model for Formative Use
11:40 am CT	1:10 pm CT	Chicago Ballroom G/H	Vendor Collaboration That Supports State Solutions
11:40 am CT	1:10 pm CT	Denver/ Houston	Research Blitz: Test Scaling, Linking, and Equating
11:40 am CT	1:10 pm CT	Los Angeles/ Miami	Modeling Test Taking Behaviors Through Process Data
1:30 pm CT	2:30 pm CT	Chicago Ballroom B/C	[Joint Session with AERA Division D] Revision of the Standards for Educational and Psychological Testing
1:30 pm CT	2:30 pm CT	Chicago Ballroom D [Recorded]	Successful NLP Approaches to Automate Scoring of NAEP's Reading Assessment
1:30 pm CT	2:30 pm CT	Chicago Ballroom E [Recorded]	Through-year Assessment Systems: Impacts on Educational Decision Making
1:30 pm CT	2:30 pm CT	Chicago Ballroom F	Investigating Measurement Invariance in Noncognitive Assessment
1:30 pm CT	2:30 pm CT	Chicago Ballroom G/H	Fairness and Equity in Assessment
1:30 pm CT	2:30 pm CT	Denver/ Houston	Demonstrations: Session 2
1:30 pm CT	2:30 pm CT	Los Angeles/ Miami	Tools and Perspectives on Assessment Literacy
2:50 pm CT	4:20 pm CT	Chicago Ballroom B/C	Improving Measures of Opportunity to Learn (OTL) to Address Systemic Inequity
2:50 pm CT	4:20 pm CT	Chicago Ballroom D [Recorded]	[SIGIMIE Session] How Can Statewide Accountability Testing Improve Student Learning?
2:50 pm CT	4:20 pm CT	Chicago Ballroom E [Recorded]	[SIGIMIE Session] Towards Culturally Relevant Assessment: Reconceiving How to Incorporate Culture into Teaching Measurement
2:50 pm CT	4:20 pm CT	Chicago Ballroom F	Simulating Large-Scale Assessment Data: Tools and Practice
2:50 pm CT	4:20 pm CT	Chicago Ballroom G/H	Tackling Through-Year Assessment Topics from a Practitioner's Point of View
2:50 pm CT	4:20 pm CT	Denver/ Houston	[SIGIMIE Session] Harmonize Tradition and Innovation: Scaling, Linking, and Equating in Technology-Enhanced Measurement
2:50 pm CT	4:20 pm CT	Los Angeles/ Miami	Computer Adaptive Testing: Models and Estimation
4:40 pm CT	6:10 pm CT	Chicago Ballroom B/C	The Impact of Pandemic on Testing Industry

Saturday, April 15

Begin Time	End Time	Room	Session Title
4:40 pm CT	6:10 pm CT	Chicago Ballroom D [Recorded]	Remembering Ron: Reflections on a Career and a Legacy
4:40 pm CT	6:10 pm CT	Chicago Ballroom E [Recorded]	Test Equity and Fairness from the Voices that Matter
4:40 pm CT	6:10 pm CT	Chicago Ballroom F	Improving Teacher Decisions in the Mathematics Classroom Through Measurement
4:40 pm CT	6:10 pm CT	Chicago Ballroom G/H	Transforming K-12 Assessments: Providing Valid Data for Instructional Decisions, Equity, and Accountability
4:40 pm CT	6:10 pm CT	Denver/ Houston	[CODIT and AERA Division D EIC Invited Session] Recruitment and Retention of Minoritized Measurement Professionals
4:40 pm CT	6:10 pm CT	Los Angeles/ Miami	Gauging Student Understanding In-The-Moment Through the Formative Assessment Process



Leveraging Measurement for Better Decisions

We are the world's learning company, driven by a mission to help people make progress in their lives through learning. We are educators, parents, research scientists, technology experts, and content specialists. Our technology-powered assessment tools, content, products, and services support millions of teachers and learners every day. Having delivered more than 100 million online tests for district, state, and national customers, we are committed to inspiring and supporting a lifelong love of learning. Because wherever learning flourishes, so do people.

Learn more at **PearsonEd.com/future-of-assessment**

Pearson

800-627-7271 | PearsonAssessments.com

001. Diagnostic Classification Models: Advanced Applications

Training Session
8:45 to 12:45 pm
Virtual: Room 1

Diagnostic classification models (DCMs) are emerging psychometrics tools that focus on providing actionable feedback from multidimensional tests. This workshop builds upon a foundational understanding of DCMs and provides a more advanced introduction to DCMs. More specifically, this workshop focuses on the structural component of DCMs, hierarchal DCMs, longitudinal DCMs, and polytomous DCMs. After completing this workshop, participants will understand the statistical structure of DCMs, be able to estimate DCMs and interpret software output, and understand how extended DCMs (hierarchical, longitudinal, polytomous) can be applied to analyze complex data sets. This session is appropriate for graduate students, researchers, and practitioners at the emerging or experienced level. Participants are expected to have only 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. Material is presented at a technical level when necessary for understanding the models and applying them responsibly. Content will mostly be delivered through lecture, and content will be reinforced using hands-on activities. The instructor will encourage attendee participation through questions and allow time for discussions among participants and the instructor.

Presenter:
Matthew James Madison, **University of Georgia**

002. Using Stan for Bayesian Psychometric Modeling (Part I)

Training Session
8:45 to 12:45 pm
Virtual: Room 2

This session will provide attendees with systematic training on Bayesian estimation of classic psychometric models as well as newly developed models using Stan, with a particular focus on helping graduate students who are searching for dissertation topics to navigate the vast body of Bayesian psychometric literature. The estimation of model parameters for common and sophisticated psychometric models will be illustrated and demonstrated using Stan. Although this workshop places a particular emphasis on IRT models, other psychometric models such as generalizability theory, classic test theory, confirmatory factor analysis, latent class models, cognitive diagnostic models, and structural equation models will also be covered. Further, the advantages and disadvantages of Stan compared to traditional Bayesian software programs such as OpenBUGS and JAGS will be discussed. This session consists of lecture, demonstration, and hands-on activities of running Stan. It is intended for intermediate and advanced graduate students, researchers, and practitioners who are interested in learning the basics and advanced topics related to parameter estimation of psychometric models using Stan.

Presenters:
Yong Luo, **NWEA**
Xin Qiao, **Southern Methodist University**

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

Training Session
1:00 to 5:00 pm
Virtual: Room 2

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

Presenters:
Seung W. Choi, **University of Texas at Austin**
Sangdon Lim, **University of Texas at Austin**

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

Training Session
1:00 to 5:00 pm
Virtual: Room 3

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

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

006. Tools For Analyzing NAEP and TIMSS Data in R Using Latent Regression

Training Session
1:00 to 5:00 pm
Virtual: Room 4

This course teaches achievement analyses with NAEP and TIMSS data using R packages EdSurvey and Dire. We introduce two analytical workflows: 1. Using the existing plausible values, and 2. A latent regression modeling of student proficiencies that is estimated directly through an MML algorithm, conditioning on student item performance and existing or new contextual variables. The second workflow allows researchers who wish to use newly constructed factors, process data, or data from other sources to get unbiased coefficient estimates in achievement analysis. In addition, new plausible values can be drawn the latent regression model for further statistical analysis. Public-use NAEP and TIMSS data files will be used for demonstration and hands-on practices via the R packages EdSurvey and Dire. Participants will learn how to perform: • data processing and manipulation, • descriptive statistics, • linear regression, • latent regression, and • plausible values generation. Participants are expected to arrive with R and RStudio installed. This course is designed for individuals who are interested in learning how to analyze NAEP and TIMSS data in R.

Presenters:
Ting Zhang, **American Institutes for Research**
Emmanuel Sikali
Paul Bailey, **American Institutes for Research**
Sinan Yavuz, **American Institutes for Research**
Blue Webb, **American Institutes for Research**

007. An Overview of Operational Psychometric Work in Real World

Training Session
8:45 to 12:45 pm
Virtual: Room 1

This training session will present an overview of the psychometric work routinely done at various testing organizations. The training session will focus on the following topics: (1) outline of operational psychometric activities across different testing companies, (2) overview and hands-on activities to review item and test analyses output, (3) overview and hands-on activities to review equating output, and (4) an overview of adaptive testing design for large-scale assessment and hands-on activities using simulation programs. We are hoping that through this training session, participants will get a glimpse of the entire operational cycle, as well as gain some understanding of the challenges and practical constraints that psychometricians face at testing organizations. It is targeted toward advanced graduate students who are majoring in psychometrics and seeking a job in a testing organization and new measurement professionals who are interested in an overview of the entire operational testing cycle. Representatives from major testing organizations (e.g., ETS, Pearson, Riverside Insights, and WestEd) will present various topics related to processes in an operational cycle.

Presenters:
Hyeon-Joo Oh, **Riverside Insights**
JongPil Kim, **Riverside Insights**
Jinghua Liu, **Pearson**
Sarah Quesen, **WestEd**
Hanwook Yoo, **Educational Testing Service**

008. Applying Data Mining Methods to Detect Test Fraud

Training Session
8:45 to 12:45 pm
Virtual: Room 2

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

Presenters:
Sarah Linnea Toton, **Caveon Test Security**
Kaiwen Man, **University of Alabama**
Yiqin Pan, **University of Florida**
Cheng Hua, **University of Alabama**

009. Using Stan for Bayesian Psychometric Modeling (Part II)

Training Session
8:45 to 12:45 pm
Virtual: Room 3

This session will provide attendees with systematic training on Bayesian estimation of classic psychometric models as well as newly developed models using Stan, with a particular focus on helping graduate students who are searching for dissertation topics to navigate the vast body of Bayesian psychometric literature. The estimation of model parameters for common and sophisticated psychometric models will be illustrated and demonstrated using Stan. Although this workshop places a particular emphasis on IRT models, other psychometric models such as generalizability theory, classic test theory, confirmatory factor analysis, latent class models, cognitive diagnostic models, and structural equation models will also be covered. Further, the advantages and disadvantages of Stan compared to traditional Bayesian software programs such as OpenBUGS and JAGS will be discussed. This session consists of lecture, demonstration, and hands-on activities of running Stan. It is intended for intermediate and advanced graduate students, researchers, and practitioners who are interested in learning the basics and advanced topics related to parameter estimation of psychometric models using Stan.

Presenters:
Yong Luo, **NWEA**
Xin Qiao, **Southern Methodist University**

010. Comparison and Integration of Generalizability Theory with Structure Equating Modeling

Paper Session
1:00 to 2:30 pm
Virtual: Room 1

Chair:
Khagendra Raj Dhakal, **King Mongkut's University of Technology Thonburi**

Participants:

Extending Bifactor Models to Account for Multiple Sources of Measurement Error
Walter Vispoel, University of Iowa; Hyeryung Lee; Hyeri Hong, California State University, Fresno

Bifactor modeling in social science research has markedly increased but continues to rely on single-occasion designs in which key sources of measurement error are inadequately modeled and/or confounded with construct variance. In research reported here, we introduce and compare three variations of multi-occasion Bayesian bifactor models that overcome these problems.

Benefits of Doing Generalizability Theory Analyses within SEM Frameworks
Hyeri Hong, California State University, Fresno; Walter Vispoel, University of Iowa; Hyeryung Lee

We demonstrate how G-theory designs can be integrated into SEM frameworks to reproduce the same variance components from ANOVA models for univariate and multivariate designs, incorporate congeneric relationships, correct for scale coarseness, account for method effects, and provide formal tests of model fit when appropriate.

Doing Multivariate Generalizability Theory Analyses within Structural Equation Modeling Frameworks
Walter Vispoel, University of Iowa; Hyeryung Lee; Hyeri Hong, California State University, Fresno

We analyzed numerous multivariate generalizability theory designs for subscale and composite scores from self-report personality measures using structural equation modeling techniques. Variance components, generalizability coefficients, and dependability coefficients for all scales within those analyses were virtually identical to those obtained from the mGENOVA package using traditional ANOVA-based procedures.

Comparing Multivariate, Bifactor, and Univariate Generalizability Theory Methods for Estimating Score Consistency
Hyeryung Lee; Walter Vispoel, University of Iowa; Hyeri Hong, California State University, Fresno

We compared generalizability and dependability coefficients and partitioning of variance for composite scores from one and two facet generalizability theory multivariate, bifactor, and univariate designs for a popular self-report measure. Score consistency indices were virtually identical for multivariate and bifactor designs but systematically higher than those for univariate designs.

Discussant:
Tony Albano, **University of California, Davis**

011. Automated Assessment of Writing and Reading Proficiency

Paper Session
1:00 to 2:30 pm
Virtual: Room 2

Chair:
Jianshen Chen, **College Board**

Participants:

Extracting Additional Information from Constructed Response Items using Latent Variable Language Modeling
Alexander Kwako; Mark Hansen, UCLA

We propose a new method for estimating latent proficiency scores from examinees' constructed response items. This approach uses human ratings as the basis from which generative latent variable language models can be trained to improve estimates. We describe the conceptual basis of the technique alongside current challenges.

Assessing the Performance of a Simple Method to Automatically Score Short-Answer Questions
Christopher Runyon, NBME; Jia Quan; Janet Mee, NBME

We present a simple method for automatically scoring short-answer questions when the expected response is a single word or phrase. The method uses string distance metrics and mixture modeling instead of complex natural language processing techniques. Results indicate the simple method works similarly well to NLP with a few exceptions.

Efficient Automated Essay Scoring using Transformer-based Active-Learning Methods
Tahereh Firoozi, University of Alberta; Hamid Mohammadi, Department of Computer Engineering, University of Amirkabir; Mark Gierl, University of Alberta

We evaluated three active learning methods than can be used to minimize the number of human scored essays required to train a modern AES system. We demonstrate that less than 5% of the original training essay sample is required to produce results that are 95% accurate using active learning methods.

Impact of MI Write Automated Writing Evaluation on Middle-Grade Writing Outcomes
Joshua Wilson, University of Delaware; Corey Palermo, Measurement Incorporated; Matthew Myers; Tania Cruz, University of Delaware; Halley Eacker, Measurement Incorporated; Jessica Coles, Measurement Incorporated; Andrew Potter, University of Delaware

This study involved a randomized controlled trial of the MI Write automated writing evaluation (AWE) system in middle-school classrooms. Results indicated no effect of MI Write on students' writing quality, writing self-efficacy, liking writing, and recursive process beliefs. Implications for the development, implementation, and consequential validity of AWE are discussed.

Sub-sequence Matching Algorithm for Improving Automated Speech Recognitions for ORF Assessment
Yihao Wang, Southern Methodist University; Eric C. Larson, Southern Methodist University; Akhito Kamata, Southern Methodist University; Joseph F. T. Nese, University of Oregon

Commercial speech recognition tools now offer linguistic features that may enable automated methods for evaluating oral reading fluency (ORF). We propose how these features can be leveraged to obtain words correct per minute (evaluated against a human scorer), and discuss modeling procedures to make ORF robust to pronunciations and phrasing.

Discussant:
Susan Lottridge, **Cambium Assessment, Inc.**

012. Models and Applications of Process Data and Eye Movement

Paper Session
1:00 to 2:30 pm
Virtual: Room 3

Chair:
Joshua Goodman, **NCCPA**

Participants:

Examining Response Processes in a Digital Performance-Based Assessment: Eye-Tracking Analyses
Yizhu Gao; Ying Cui, University of Alberta; Dongran Wang, Tobii Electronic Technology Suzhou Co., Ltd; Bin Zheng, University of Alberta

In this study, we used eye-tracking techniques to make detailed observations of item response processes in a digital performance-based assessment for measuring data evaluation competencies. By analyzing eye-tracking data of 34 adult respondents, we identified and profiled eye-movement patterns and cognitive processes associated with item performance.

Using Graphical Social Network to Model Eye Movements in Spatial Reasoning Problems
Kaiwen Man, University of Alabama; Jake Feiler, University of Alabama; Joni Lakin, University of Alabama

Eye tracking has drawn much attention in educational assessment to understand students' cognitive processes during problem solving. To explore individual differences in strategy use while solving spatial problems, this study proposes an innovative approach using item-level social networks to visualize and model eye movement sequential patterns.

Student Engagement Study of a Statewide Summative Assessment
Marc W. Julian, DRC; Litong Zhang, DRC; Xiao Zhang, DRC; Daisy Ye, DRC

Item response times were used as an indicator of examinee engagement during a statewide test administration. Rapid guessing was flagged at item and student level. To evaluate the impact of disengagement, real scores were compared with corrected scores for rapid guessing, the two scores were found to be systemically consistent.

Teachers' Digital Classroom Assessment Literacy: Exploring Behavioral Indicators on an Online Platform
Jinnie Choi, Savvas Learning Company

While digital assessments offer benefits for teachers, how teachers use them for classroom assessment purposes may or may not support learning. This study aims to support impactful use of digital assessments by describing components of digital classroom assessment literacy and exploring behavioral indicators in an online teaching and learning environment.

Discussant:
Peter Halpin, **UNC-Chapel Hill**

013. The Impact of COVID-19 and Learning Recovery

Paper Session
1:00 to 2:30 pm
Virtual: Room 4

Chair:
Olga Kunina-Habenicht

Participants:

Viewing CAT Assessments' Validity Issues Through Response Times: Pre-Post Pandemic
Chalie Patarapichayatham, NWEA; Victoria Locke, Istation
This study aims to investigate the students' response time before and after the pandemic to better understand the impact of the pandemic on students' behavior and validity issues in CAT assessments.

Evaluation of the Pandemic Impacts on Student Achievement
Yuan Hong; Stephan Ahadi, Cambium Assessment, Inc.
We propose two analysis strategies designed to control for changes in the tested population in the statewide achievement assessments sto examine pandemic-related impacts on student achievement and to identify the demographic groups that might have been differentially impacted.

COVID-19 Pandemic Impact on Achievement Considering Item Cognitive Difficulties and Cognitive Reasoning
Sharon Frey, Riverside Insights; Sid Sharairi, Riverside Insights; JongPil Kim, Riverside Insights
This study compares student's performance on achievement estimated pre-pandemic vs. post-pandemic using multiple years of operational data for the same grade cohort. The magnitude of the differences between pre- and post-pandemic will be compared and evaluated with respect to the item cognitive difficulty levels and cognitive reasoning scores assessed.

COVID-19 Learning Recovery Signal
Chalie Patarapichayatham, NWEA; Victoria Locke, Istation
This study aims to investigate the learning disruption due to the COVID-19 pandemic, whether there is a learning recovery signal, and how much students lag from their pre-pandemic levels. A piecewise growth model analyzes longitudinal CAT assessment data across four school years.

Examining Consistency of District Performance between Administrations in the Context of COVID-19
Shuqin Tao, Cambium Assessment, Inc
This paper reports an investigation of consistency in district level achievement from pre- to post-pandemic, as well as consistency in achievement during post-pandemic recovery. Given changes are expected due to the pandemic, weighted regression is used to evaluate consistency with respect to greater or lesser than expected changes in performance.

Discussant:
Susan Lyons, **Lyons Assessment Consulting**

014. Approaches for Evaluating and Reporting Strength of Validity Evidence for Assessments

Coordinated Paper Session
1:00 to 2:30 pm
Virtual: Room 5

Validity is the most fundamental consideration in developing and evaluating tests (AERA et al., 2014). While the literature describes considerations for evaluating strength of validity evidence (e.g., AERA et al., 2014; Chapelle, 2021; Cizek, 2020; Kane, 2013), there are few examples providing methods and practical steps to guide practitioners in evaluating, synthesizing and reporting validity evidence. The participants in this session will describe approaches they are using or considering in this regard. Staff from ATLAS at the University of Kansas will describe their efforts to evaluate strength of validity evidence borrowing concepts, methods, and tools used in contribution analysis (Mayne, 2012), a theory-based method of program evaluation. Staff from Cognia will describe an analytic approach to evaluate score interpretation and use statements (SIUs) on three dimensions: relevance, completeness, and overall support for the claim. Staff from the WIDA Consortium will describe their work to develop an Assessment Use Argument (AUA) along with resources to communicate validation models more effectively to various stakeholders. Finally, Michelle Croft will describe how to communicate a validity argument in an accessible way to a non-measurement audience, using methods from the legal field. Scott Marion, a national leader in assessment, will serve as session discussant.

Session Organizer:
Jennifer Kobrin, **ATLAS: University of Kansas**

Chair:
Jennifer Kobrin, **ATLAS: University of Kansas**

Participants:

Using Contribution Analysis to Evaluate the Validity Argument for an Assessment System

Jennifer Kobrin, ATLAS: University of Kansas; Amy Clark, ATLAS: University of Kansas; William Jacob Thompson, University of Kansas

Contribution analysis (Mayne, 2012), a theory-based method of program evaluation, may lend itself to evaluating strength of an assessment's validity argument. We will discuss our experiences adapting methods and tools from contribution analysis to evaluate the strength of validity evidence, sharing successes and challenges in this endeavor.

An Analytic Approach to Validity Argumentation for Test Scores

Steve Ferrara, HumRRO; Louis Roussos, Cognia; Qi Qin, Gwinnett County Public Schools

In this paper, we will provide examples to illustrate evaluations of evidence for several claims about test scores and describe the benefits and drawbacks of taking an analytic, multi-dimensional approach to validity argumentation.

Creating Relevant and Accessible Descriptions of Validity Arguments

Michelle Croft

This paper will provide an overview of how to communicate a validity argument to a non-measurement audience through the selection of the strongest, most relevant pieces of validity evidence and then communicating it in an accessible way, using methods from the legal field.

Discussant:

Scott Marion, **National Center for the Improvement of Educational Assessment**

015. Platforms and Strategies to Enhance Learning

Paper Session

2:45 to 3:45 pm

Virtual: Room 1

Chair:

Anna Zilberberg

Participants:

A Collaborative Problem-Solving Platform to Measure Understanding on a Mathematics Learning Progression

Jessica Andrews Todd, Educational Testing Service; Edith Aurora Graf, Educational Testing Service; Wallace Nascimento, University of Florida

Collaborative problem-solving (CPS) is a valued 21st-century skill, and a learning progression (LP) has the potential to inform instructional decisions. We describe results from a usability study testing an online CPS platform currently under development. The platform measures student progress with respect to a mathematics LP.

Adaptive Learning Reward Function Improvement Considering Learning Efficiency

Tongxin Zhang; Canxi Cao, Beijing Normal University; Tao Xin, Beijing Normal University

Adaptive learning reward function, which reflects the goal of learning efficiency, could be improved regarding the factors that are learning duration cost on materials and learners' a priori knowledge. This study conducted a simulation to exam the reinforcement learning recommendation strategies and gave interpretable expressions.

Visualizing Assessment Data for Personalized Learning Using the Interaction Map Approach

Eric M Ho; Minjeong Jeon, UCLA

Personalized learning, which can raise student achievement, requires understanding the competencies of students. Visualizations can provide this understanding. We propose visualizations based on the latent space item response model proposed by Jeon et al. (2021) that can convey individual student competencies and promote personalized learning.

Discussant:

Sonya Powers, **Edmentum, Inc.**

016. Research Blitz: Advances in Item Response Theory

Research Blitz Session
2:45 to 3:45 pm
Virtual: Room 2

Chair:
Qi Diao, ETS

Participants:

Assessing the Performance of the Urnings Algorithm with Single-Sitting Tests
Lanrong Li, Amplify Education, Inc.; John Stewart, Amplify Education; Reginald Ziedzor, University of Southern Illinois Carbondale

This study examined the performance of the Urnings algorithm, a recently proposed method for tracking the change in parameters over time, with one-sitting tests. The results showed that besides sample size and test length, item order, initial ratings for persons, and missing data all affected the parameter estimates.

Exploring Item Position Effects in PISA using Multilevel Three Parameter IRT models
Minsung Kim, ACT, Inc.; NooRee Huh, ACT, Inc.

A multilevel approach using a three-parameter logistic (3PL) item response theory was developed to investigate the sources of item position effects to estimate item parameters in different countries. The analyses results showed that item position effects on estimating item parameters were bigger in countries with higher country variable mean scores.

The Dynamic Rasch Model and Thurstone's Learning Curve: An Extension and Illustration
Wanchen Chang, Cambium Assessment; Seyfullah Tingir, Amplify Education; Guoguo Zheng, Amplify Education

This study proposes an extension to the dynamic Rasch model based on Thurstone's learning curve. We evaluated this extension using data from an assessment for learning system. Our extension produced final ability estimates that were more correlated with scores on a posttest, compared to the linear version of the model.

A Comparison of Symmetric and Asymmetric Item Response Theory Models
Xing Chen, Fordham University; Leah Feuerstahler, Fordham University

This study compares the fit of symmetric and asymmetric item response models in real and simulated data. In real data, the residual heteroscedasticity model yielded the best fit and lowest intra-item parameter correlations. In simulations, fitting data to the correct model inconsistently yielded the closest fit or lowest information criteria.

A Multi-Unidimensional Pairwise-Preference Model for RANK Response Format Data
Wenqing Zhang, East China Normal University (Intern, Beijing Insight Online Management Consulting Co., Ltd); Chanjin Zheng, East China Normal University; Juan Liu, Beijing Insight Online Management Consulting Co., Ltd; Yalin Li, Beijing Insight Online Management Consulting Co., Ltd; Xu Lian, Beijing Insight Online Management Consulting Co., Ltd

In large-scale testing with high dimensions, the current forced-choice models are prone to non-convergence

In large-scale testing with high dimensions, the current forced-choice models are prone to non-convergence and inefficiency. In this study, a 2PL-RANK model for RANK response format was suggested, with stochastic EM algorithm to estimate parameters. The simulation and empirical findings demonstrate the effectiveness and efficiency of the 2PL-RANK model.

How Many Is Enough to Calibrate Single or Mixed IRT Models?
Hye-Jeong Choi, Human Resources Research Organization; Dipendra Subedi, Pearson; Yufeng Berry, Minnesota Department of Education; Meng Fan; Gerald Griph, Pearson; Changjiang Wang, Pearson; Yvette M Nemeth, HumRRO

This study intends to investigate the effects of sample size on single and mixed IRT model parameter calibration. We will also compare the performance of three software packages (PROC IRT, IRTPRO, and mirt). A simulation study will be conducted. An empirical data analysis will be presented with practical suggestions.

017. Standard Setting and Proficiency Level Descriptors

Paper Session
2:45 to 3:45 pm
Virtual: Room 3

Chair:
Yizhu Gao

Participants:

Writing Achievement Level Descriptors to Maximize Interpretability
Robert J. Cook, Cognia, Inc.

Valid interpretation of test scores requires careful and precise articulation of the interpretations that test scores are meant to take on. This paper demonstrates how the most commonly applied framework for achievement level descriptor writing can threaten valid interpretation and offers a new framework that preserves interpretability by design.

Using Locally-Derived Cut Scores to Improve Universal Screening for All Students
Quentin Ulysses Adrian Love, WestEd

Early literacy universal screening is high stakes for individual students. But current national cut scores do not function equally well for all students. Given multiple years of statewide assessment data, we investigate whether using locally-derived cut scores improves the diagnostic accuracy of reading screeners. Results suggest improvement for all students.

Well-Informed Cut (WIC) Standard Setting
Richard Melvin Luecht, University of North Carolina at Greensboro; Chad W. Buckendahl, ACS Ventures, LLC; Joshua Goodman, NCCPA; Leslie Keng, Center for Assessment

The Well-Informed Cut (WIC) standard setting method leverages technology and simplifies what the panelists do to set/modify their recommended cut(s). The method flips the process and actively engages panelists in understanding the impact of their decisions on examinees and items. An empirical study and operational R/Shiny software are exhibited.

Discussant:
Adam E Wyse, **Renaissance Learning**

018. GSIC Virtual eBoard Session

Graduate Electronic Board Session
2:45 to 3:45 pm
Virtual: Room 4

Participants:

MCMC Convergence Diagnostics in the DINA Model and the Bi-factor IRT Model
Sunbeom Kwon; Susu Zhang, University of Illinois at Urbana-Champaign; Hans Friedrich Koehn

The objective of this study is to compare the performance of two diagnostic tools for evaluating the MCMC convergence in Bayesian estimation of latent variable models. The commonly used Gelman-Rubin diagnostic was found to prematurely flag convergence for both discrete and continuous latent variable models considered here.

Mining Textual Features of Questions to Predict Item Parameters
RBin Tan, University of Alberta; Okan Bulut, University of Alberta; Guher Gorgun, University of Alberta; Tarid Wongvorachan, University of Alberta

Parameter calibration in educational assessments often requires a large sample, which involves significant test administration costs. This study uses machine learning to predict item parameters based on the textual features of questions. Results indicate that the textual features may be useful in item parameters without test administration.

Systematic Review of the Use of Process Data in Large-scale Assessments
Surina He; Ying Cui, University of Alberta.

This ongoing study conducted a systematic review of the use of process data. We found that the number of studies on the use of process data in large-scale assessment has been on the rise annually. And 2021 is the year with the most publications until now.

Validation as Evaluating (Un-)desired Effects: Insights from Cross-Classified Mixed Effects Model
Xuejun Ryan Ji, The University of British Columbia; Amery Wu, University of British Columbia

This study aims to 1) showcase how validation can be undertaken as an exercise of identifying and explaining sources of desired and undesired effects (or score variations), 2) elaborate the fruitfulness of cross-classified mixed effects model as a validation tool.

The Mantel-Haenszel and Logistic Regression DIF Methods for Formative Digital Assessment Items
Lissette Tolentino

The use of formative digital assessment items in virtual learning environments is becoming very common. However, little is known about their psychometric properties in relation to test fairness. This study investigates the efficacy of differential item functioning methods as it relates to these items to support test equity and fairness.

A Comparison of Sampling Methods for Imbalanced Classifications in Educational
Tarid Wongvorachan, University of Alberta; Surina He; Ka Wing Lai; Okan Bulut, University of Alberta

Data imbalance reduces prediction accuracy in classification models for educational dataset since algorithms often favor the majority class. We compare several sampling techniques to handle the data imbalance problem. Random oversampling for moderately imbalanced data and hybrid resampling for extremely imbalanced data seem to work very well.

019. Factor Analysis Model Fit

Paper Session
4:00 to 5:00 pm
Virtual: Room 1

Chair:
Martha McCall, **McKinsey & Company**

Participants:

Discrepancies between CFI and RMSEA
Menglin Xu; Paul De Boeck, OSU

Goodness of fit indices for CFA has been lively discussed, and the potential discrepancies between CFI and RMSEA in evaluating models have been called to attention. This study uncovers the mystery through theoretical analysis and simulation study. Hybrid nature of CFI was identified. Implications are discussed.

Investigating Model Fit Indices in Multiple-group Confirmatory Factor Analysis with Ordinal Data
Ning Jiang; Christine DiStefano, University of South Carolina; Jie Chen, Measurement Incorporated

This study evaluated the performance of CFI, RMSEA, and SRMR when measurement invariance is tested using a multiple-group CFA with ordinal data. A Monte Carlo simulation study was conducted to examine the sampling variability of fit indices. Cutoff values for various levels of invariance were proposed.

Comparing Accuracy of Parallel Analysis and Fit Statistics in EFA
Hyunjung Lee; Heining Cham, Fordham University

This study aims to compare the parallel analysis to the performance of fit indices in EFA. The Monte Carlo simulation study was conducted with ordered categorical items. The results indicate that the parallel analysis and RMSEA performed well in most conditions, followed by TLI and then by CFI.

Discussant:
Sanford Student, **University of Colorado Boulder**

020. Demonstrations: Software and Training Module

Demonstration Session
4:00 to 5:00 pm
Virtual: Room 2

Chair:
Yu-Lan Su, **Ascend Learning**

Participants:

Applied Diagnostic Classification Modeling with R Package measr
W. Jake Thompson, University of Kansas

Diagnostic assessments provide reliable and actionable results with shorter test lengths. However, these methods are not often used in applied research due to in part to limited and inaccessible software. In this presentation we describe a new and free software, measr, that can easily estimate and evaluate diagnostic models.

simpleIPD: An R Tool for Evaluating Common Item Parameter Drift
Daniel Yangsup Lee, College Board; Youngkoung Kim, College Board; Tim Moses, College Board

SimpleIPD is a tool that can help evaluate the item parameter drift of common items between different testing occasions. The tool presents several drift statistics that may be helpful for assessing the impact of retaining and removing anchor items for use of scaling in nonequivalent groups with anchor test designs.

An Exploration of Developing a Training Module for Negotiation Skills
Yuan Wang, ETS; Jennifer Lentini, Educational Testing Service (ETS); Zhitong Yang; Emily Kerzabi, Educational Testing Service; Salenah Cartier, ETS; Guangming Ling, Educational Testing Service; Michelle Martin-Raugh, University of Texas Arlington

In this innovation demonstration we will share a training module developed to strengthen negotiation skills for those entering the workforce. The module includes multimedia training and multiple opportunities for participants to practice negotiation techniques with a partner facilitated by an automated intelligent tutor.

021. The Roles of Distractors in Developing Digital Assessments Within Assessment Engineering Frameworks

Coordinated Paper Session
4:00 to 5:00 pm
Virtual: Room 3

Modern psychometricians and assessment service providers increasingly find themselves facing a new paradigm, Assessment Engineering (AE) in which assessment production and management processes can no longer be analogous/manual, but instead can be digitalized/automated by Knowledge Engineering technologies. Although nearly 50 years have passed since the introduction of various AE methodologies (e.g., Evidence-centered Design [ECD], Automatic Item Generation [AIG], etc.) and studies that present a rosy future of AE. Practical applications/implementations of these methods are still expensive, slow, and narrow. This is because sufficiently detailed strategies and methodologies to effectively/effectively implement AE methods in the field have not yet been developed/introduced. The concept, role, and usage of distractor in traditional assessment development are extended to Distractor Modeling (DM) in AE, and the roles/usages of DM is further extended in various AE developments (e.g., ontology model or formative assessment model **developments). The topics of this session** are intended to present innovative strategies for practical implementations of AE centered around DM. This session is a collection of ongoing studies on what role Distractor Modeling plays in AE and how to use it to develop digital assessments more effectively/efficiently - especially focused on sustainability and innovation.

Session Organizer:
Jaehwa Choi, **George Washington University**

Participants:
Reverse Assessment Knowledge Engineering: The Role of Distractors in Reverse Engineering within Assessment Engineering Framework
Shonai Someshwar, UNC Greensboro; Eunji Lee, George Washington University.
Implications of Distractor Model Specifications onto Assessment Quality in Automatic Item Generation
Sunhyoung Lee, University of Nebraska-Lincoln
Smart Distractor Modeling for Vocabulary Test within Automatic Item Generation Framework
Kyuseol Oh, Geunhwa Girls High School

022. Virtual eBoard Session Electronic Board Session

4:00 to 5:00 pm
Virtual: Room 4

Participants:
Time-on-Task from Log and Eye Movement Data: Commonalities and Differences
Tobias Deribo, DIPF | Leibniz Institute for Research and Information in Education; Ulf Kroehne; Carolin Hahnel, DIPF | Leibniz Institute for Research and Information in Education, Centre for International Student; Frank Goldhammer, DIPF | Leibniz Institute for Research and Information in Education, ZIB

Time-on-task can be helpful for multiple psychometric applications. However, when a multiple-item-per-page design is used, relating time-on-task to a specific item may be difficult. Therefore, we investigated this problem by comparing time-on-task measures based on eye movement and log data. Overall, the results indicate only negligible differences between measures.

The Effect of Mixed-Format Item Pools on Computerized Adaptive Testing
Lucia Liu, Ascend Learning; Ye Lin, Ascend Learning

Two factors are considered for a mixed-format CAT – the proportion of polytomous items and the maximum number of item score points. The findings indicate that the proportion of polytomous items is more critical for CAT's performance. A balanced combination of dichotomous and polytomous items achieves the best performance.

Random Search Algorithm to Identify Response Time Thresholds for Rapid Guessing
Tarid Wongvorachan, University of Alberta; Okan Bulut, University of Alberta; Guher Gorgun, University of Alberta; Bin Tan, University of Alberta

A popular approach for identifying rapid guesses is the threshold-based method. Typically, an arbitrary threshold is defined and applied to all items. This study introduces a novel data-driven approach for identifying rapid guesses. The proposed approach outperformed the previous method suggesting that thresholds should be optimized for each item.

Bayesian Comparison of Growth Mixture Models
Xingyao Xiao, XXY; Sophia Rabe-Hesketh, University of California, Berkeley; Feng Ji

In Growth Mixture Models selection, finding the number of latent trajectory classes is important and challenging. Researchers often use ad-hoc approaches for convenience, but many of these methods perform poorly. This paper shows that Bayesian model selection via marginal likelihood is a rigorous approach that performs well in various circumstances.

Measurement Invariance Across Immigrant and Non-Immigrant Populations on PISA Non-Cognitive Scales
Maritza Casas, University of Massachusetts Amherst; Stephen G Sireci, University of Massachusetts, Amherst.

In this study we compared multigroup confirmatory factor analysis and the alignment optimization method to evaluate the invariance of the bullying and sense of belonging at school PISA scales across groups of students defined by immigrant status. Alignment optimization produced more useful results revealing the invariance properties of the scales.

M-Estimation of Principal Effects Using Principal Scores and OLS
Adam C Sales, Worcester Polytechnic Institute; Kirk P Vanacore, Worcester Polytechnic Institute; Erin Ottmar, Worcester Polytechnic Institute

We present a novel--if simple--approach to estimating principal effects under one-way noncompliance based on stacked estimating equations from logistic and OLS regressions. We give general conditions under which estimators and standard errors are consistent, present simulation results showing good finite-sample performance, and demonstrate the method with educational RCT data.

Impact of Rater Effects on Classification Consistency and Accuracy in Performance-Based Assessments
Daniel Edi, Pearson Assessments and Qualifications

This study simulated two levels of four types of rater effects and assessed their impact on examinee classifications under the MFRM framework. Results indicated that up to 10% and 50% of examinees were inconsistently and inaccurately classified, respectively, due to the presence of examiners exhibiting rater effects.

Impact of College Entrance Exam Mandate on College Readiness and Enrollment
Burhan Ogut, American Institutes for Research; Yusuf Canbolat, Indiana University Bloomington

This study examines the effect of the college entrance exam mandate on college readiness and enrollment. The study uses data from the High School Longitudinal Study of 2009 and employs a multilevel propensity score matching approach, logistics regression, and hierarchical linear regression. Results indicate that taking the college entrance exam (ACT or SAT) improves college readiness and enrollment.

Predicting IRT 3PL Parameters of Reading Comprehension Items
Dmitry I. Belov, Law School Admission Council; Anna Topczewski, Law School Admission Council; Aaron McVay, Law School Admission Council

A neural network model to predict IRT 3PL parameters of passage based multiple-choice reading comprehension items from the Law School Admission Test was developed based on multiple features automatically extracted from the passage and item text. Major stages of the development (design, training, and validation) are demonstrated.

Changes in Research Topics in High School Credit System Using Topic Modeling
Eunjeong Jeon, Ewha Womans University; Youn-Jeng Choi, Ewha Womans University; Ji-Hye Kim, Korean Educational Development Institute

The purpose of this study is to investigate changes in topics over time by analyzing academic papers related to the high school credit system. The analysis will apply latent Dirichlet allocation using the Topicmodels R package. Implications for the policy direction can be obtained through this study.

023. Presenting from Three Continents on Three Topics: Collaborative Problem Solving, Linked Scores, and Propensity Score Estimation

Paper Session

5:15 to 6:45 pm

Virtual: Room 1

Chair:

Robert Thomas Furter, **Physician Assistant Education Association**

Participants:

Exploring The Relationships Between Small Group and Individuals in Collaborative Problem Solving
Nafisa Awwal, University of Melbourne; Mark Wilson, Berkeley School of Education, UC Berkeley; Zhonghua Zhang, University of Melbourne

The authors used a process-based multidimensional framework to assess collaboration in groups and individuals interacting in computer-based problems. The Rasch model is applied on interactional process data to help interpret empirical aspects of the collaboration. The results of the empirical data are further investigated to explore the relationships between small groups and individuals in them during collaborative problem solving.

Can We Use the Linked Scores as Predictions of Individual Scores?
Yoshikazu Sato, Admission Center, Kyushu University; Tadashi Shibayama, Tohoku University

We developed a method to quantitatively evaluate the predictive ability of linked scores. Using this method, the examples of concordance and equating in the Japan Law School Admission Test were used to examine the prediction accuracy and error properties of the linked scores.

Propensity Score Estimation with Multi-Layer Neural Networks
Igor Migunov, Fordham University; Heining Cham, Fordham University

A simulation study was conducted to test the utility of using multi-layered neural networks in estimating the propensity scores. Results supported its utility in the scenarios when both the outcome and propensity score models are non-linear. However, in the case of low exposure prevalence adding each hidden layer must be done with caution.

Discussant:

Okan Bulut, University of Alberta

024. Identification of Low-Effort Responses and Measurement of Digital Literacy

Paper Session
8:45 to 10:15 am

Virtual: Room 1

Chair:
Samuel Haring, **ACT**

Participants:

Establishing Response Time Threshold to Identify Low-effort Examinees through Latent Profile Analysis
Ismail Cuhadar, Ministry of Education, Turkey; Meltem Yumsek-Akbaba, Ministry of Education, Turkey

Time-on-task can be helpful for multiple psychometric applications. However, when a multiple-item-per-page design is used, relating time-on-task to a specific item may be difficult. Therefore, we investigated this problem by comparing time-on-task measures based on eye movement and log data. Overall, the results indicate only negligible differences between measures.

Modeling Partial Inattentive Responses in Mixed-Format Scales
Kuan Yu Jin; Ming Ming Chiu, Education University of Hong Kong

Surveys' reverse-coded questions reveal inattentive respondents but they need not answer questions consistently. Hence, we expanded the mixture model for inattentive responses to show how they can cause bias and reduce test reliability and accuracy in slope and intercept parameters with an empirical analysis.

Measuring Digital Literacy via a Performance-Based Assessment: A Longitudinal Cohort Study
Qianqian Pan, National Institute of Education, Nanyang Technological University; Qianru Liang, The University of Hong Kong; Nancy Law, University of Hong Kong; Frank Reichert, The University of Hong Kong; Jimmy de la Torre, University of Hong Kong

This paper describes the development of a performance-based digital literacy assessment for measuring DL development from late childhood to early adulthood via a longitudinal cohort study design. Using data from three age cohorts at two time points over two years, the psychometric properties of the assessment were examined.

Discussant:
Kimberly Swygert, **National Board of Medical Examiners**

025. Multidimensionality and Adaptive Learning Modeling

Paper Session
8:45 to 10:15 am
Virtual: Room 2

Chair:
Chris Patterson, **James Madison University**

Participants:

Detecting Multidimensional DIF in Polytomous Items with IRT Methods and Estimation Approaches
Güler Yavuz Temel, Hamburg University

The purpose of this study was to investigate the performance of the multidimensional DIF with IRT based methods in a simulation study and with PISA 2018 student questionnaire. The results showed that when sample size and DIF magnitude was large, DIF were correctly identified with IRT based approaches and estimations.

A Pre-Rule Check for The Conditional Multidimensional Sequential Probability Ratio Test
Bo Sien Hwu; Cheng Te Chen, National Tsing-Hua University, Taiwan; Ching-Lin Shih, National Sun Yat-sen University

The information shared between dimensions was incorporated into the SPRT by Liu et al. (2021) and found the classification efficiency was improved except for the condition when the cutoffs were set to the population means. This study proposed a pre-rule check to overcome this shortcoming while maintaining its classification accuracy.

Research on Measurement Model of Adaptive Learning System Based on XGBoost Algorithm
Chang Nie, Beijing Normal University; Tao Xin, Beijing Normal University

In this study, a measurement model of adaptive learning system based on XGBoost algorithm was constructed. The simulation results show that the XGBoost-based measurement model has higher accuracy than DINA, especially in case of short test length; moreover, the adaptive learning system applying this new measurement model is more effective.

Discussant:
Ru Lu, **Educational Testing Service**

026. Test Equating and Linking Challenges and New Methodology

Paper Session
8:45 to 10:15 am
Virtual: Room 3

Chair:
Kari Hodge

Participants:

Comparison of Equating Methods when DIF is Present in Common Items
Gamze Kartal, University of Illinois at Urbana-Champaign

Scores on test forms administered at different times to different examinees can be used interchangeably if test equating is used. Investigating the performances of equating methods when DIF is present in common items will aid in the selection of the most robust equating method to achieve validity of equating results.

Equating Subscores and Overall Score in Multidimensional Test Data
Aysenur Erdemir; Won-Chan Lee, University of Iowa

Tests may be inherently multidimensional due to the intended content or construct structure of the tests. The primary purpose of the present research is the present an observed-score equating procedure for overall score and subscores using higher order item response theory model under a random groups design. The results will be reported.

Empirical Ensemble Equating under the NEAT Design Inspired by Machine Learning Ideology
Zhehan Jiang; Yuting Han, Peking University; Lingling Xu, Peking University; Jinying Ouyang; Jihong Zhang, University of Iowa; Ren Liu, University of California, MERC; Dexin Shi, University of South Carolina

Different statistical techniques used in the Non-Equivalent groups with Anchor Test (NEAT) tasks tend to yield inconsistent performance across equating settings and/or score ranges. In order to take and combine the advantage of equating techniques in various score intervals, this study proposes an empirical ensemble equating (3E) approach that collectively selects, adopts, weighs, and combines outputs from different sources. The ensemble idea was demonstrated and tailored to the NEAT equating. A simulation study showed that the 3E approach is valuable to practical inquiries.

Scaled Score Change of Short Forms Derived from Long Form
Rui Gao, ETS

The study examines when a regular test form (long form) was divided into several shortened forms (short form), whether items would perform differently and how the scaled score of the short forms would change.

Discussant:
Anna Topczewski, **Law School Admission Council**

027. Cognitive Diagnosis Models and Practice

Paper Session
8:45 to 10:15 am
Virtual: Room 4

Chair:
Soo Ingrisone, **Pearson**

Participants:

Cognitive Diagnosis Model for Item and Person Random Effects
Youn Seon Lim, University of Cincinnati

The key assumption underlying cognitive diagnosis models is the local independence of item responses. In educational settings, this assumption is often violated, which leads to grave consequences for the test validity and reliability. To solve this issue, this study proposes a cognitive diagnosis model incorporating item and person random effects.

Using ROC statistics to assess model fit for diagnostic classification models
Cheng Hua, University of Alabama; Wenchao Ma, University of Alabama

This study examines the feasibility of using the Receiver Operating Characteristic (ROC) statistics including the area under the curve (AUC) and F1 score to assess the model-data fit for diagnostic classification models at either item or test level by comparing the correct model selection rate with traditional DCM fit statistics.

Assessing Item-Level Fit for the Sequential Process Model
Pablo Nájera, Autonomous University of Madrid; Wenchao Ma, University of Alabama; Miguel A. Sorrel, Universidad Autónoma de Madrid; Francisco J. Abad, Universidad Autonoma de Madrid

The sequential process model is a cognitive diagnostic model that can accommodate graded responses. The present study aims to explore the performance of several item-level fit statistics to detect the presence of model and Q-matrix misspecifications. Practical guidelines are given to facilitate the detection and remedy of misfitted items.

Theory for Building Proper and Useful Distractors for Cognitive Diagnosis
Hans Friedrich Koehn; Chia-Yi Chiu, University of Minnesota; Yu Wang, University of Minnesota, Twin Cities
When multiple-choice items are used for cognitive diagnosis, it is crucial to make sure that the distractors are useful and proper so that they can indeed improve the correct classification rates and avoid misclassifications. Two criteria are proposed **in the study to identify such distractors.**

Detecting Misconceptions: A Quest to Convert Imperfect Information into Learning Opportunities
Jennifer L. Lewis, University of Massachusetts Amherst
This study explores the reliability and accuracy of detecting misconceptions under the diagnostic concept inventory framework (Bradshaw et al., 2022) when the response data is derived from discrete option multiple-choice (DOMC) SmartItems (Foster, 2016). The results will be used to empirically investigate the appropriateness of DOMC SmartItems for detecting misconceptions.

Discussant:
Jinsong Chen, **The University of Hong Kong**

028. Methods and Applications of Survey Research and Noncognitive Assessment

Paper Session
8:45 to 10:15 am
Virtual: Room 5

Chair:
Shuqin Tao, **Cambium Assessment, Inc.**

Participants:

Negative Wording Effects within Variations of Conventional, ESEM, and BSEM Factor Models
Hyeri Hong, California State University, Fresno; Walter Vispoel, University of Iowa
Method effects are common in self-report measures and can cause misinterpretation of results if not properly controlled. We investigated effects of controls for negative item wording within variations of multi-factor, hierarchical, and bifactor models and estimation procedures. Results revealed that such controls were important for some models but not others.

Bias in Job Analysis Survey Ratings Attributed to Order Effects
Rebecca Berenbon; Bridget McHugh, Center on Education and Training for Employment; Abena Anyidoho, The Ohio State University
We examined a job analysis survey for order effects. We observed an association between average ratings for survey blocks and the order in which the respondent saw the survey block. This was observed for both completers and non-completers. The results highlight the importance of using randomization to mitigate order effects.

Exploring Careless Responding in Survey Research with Mokken Scaling: An Iterative Approach
Stefanie A. Wind, University of Alabama; Yurou Wang, University of Alabama
We used real and simulated data to explore the alignment between careless responding in survey research and response quality indicators from the nonparametric Mokken Scale Analysis (MSA) approach to item response theory (IRT) using standalone and sequential procedures. MSA indicators reflect carelessness patterns and help analysts interpret survey results.

Psychometric Properties of Selweb and its Use in Measuring Student Social-Emotional Learning
Chun-Wei Huang, WestEd; Linlin Li, WestEd; Kylie Flynn, WestEd
This study aims to establish the psychometric properties of SELweb used in an efficacy study. An IRT model was used to score students' responses, followed by a validity and reliability study. The findings indicate that SELweb possesses desired psychometric properties. Thus, one can interpret the impact results with confidence.

¡Leamos! Establishing Technical Adequacy Evidence for a Spanish Literacy Screening Assessment
Deni Basaraba, Amplify Education; Lanrong Li, Amplify Education; Sandra Pappas, Amplify Education; Norma Medina Morales, Amplify Education; Danielle Damico, Amplify Education
Successfully implementing Spanish-English bilingual programs requires technically-adequate measures in both languages. Despite the rapid increase in dual-language programs, reliability and validity evidence for Spanish assessments is limited. This paper will describe the development process of an authentic universal screener of Spanish literacy for Grades K-6, including evidence of technical adequacy.

Discussant:
Michael C. Rodriguez, **University of Minnesota**

029. Using Eye Movement and Natural Language Processing to Inform Various Decisions

Paper Session
10:30 to 12:00 pm
Virtual: Room 1

Chair:
Katrina Borowiec, **Boston College**

Participants:

Using Eye Movement to Study Distinct Response Style of the Students
Ayfer Sayin; Ergün Cihat Çorbacı, Gazi University; Mehmet Fatih Doğuyurt, Gazi University
This paper uses eye movement data to study distinct response styles of 48 subjects who responded to 10 items from the BIG 5 Questionnaire in a lab setting. Initial findings are presented next, along with a preliminary discussion about potential implications for future R&D efforts.

Identifying Early Warning Indicators for High School Dropouts
Surina He; Tarid Wongvorachan, University of Alberta; Okan Bulut, University of Alberta; Ka Wing Lai
The current study identified early warning indicators for high school dropouts based on 22,612 samples from HSLS 2009 dataset. Results showed that the deep neural network achieved 70% accuracy. In addition, 9th grade GPA, significant others' expectations, sense of school belonging, and school climate were found as important actionable predictors.

An NLP Approach to Evaluating Construct Representation and Dimensionality of Item Pools
Yi-Chen Chiang, NABP; Michael R Peabody, National Association of Boards of Pharmacy
The study explores a topic modeling approach to evaluating construct representation and dimensionality of item pools. Real data from a licensure examination are used to conduct the analysis. The topics extracted from each model and its alignment with the original domain conceptualization are discussed.

Leveraging Natural Language Processing to Augment Practice Analysis
Bharati Belwalkar, AIR; Christina Curnow, AIR; Luke Patterson, AIR; Matthew Schultz, AICPA; Sandeep Shetty, AIR; Joshua Stopek, AICPA
This paper discusses how natural language processing (NLP) approaches can augment practice analysis in evaluating the state of a profession. NLP approaches were employed to extract skill information from a large-scale job posting dataset, explore trends in skill demand and derive speculative emergent themes in the practice of accounting.

Comparisons of Feature Selection Techniques in Machine Learning Approaches for Collusion
Soo Ingrisone, Pearson; James Ingrisone, Pearson VUE
Feature selection is challenging for building machine learning models. This study compares the selected features and provides guidance in choosing optimal feature selection strategies for detecting aberrant examinees. The performance of three wrapper models by three classifiers in machine learning approaches under nine different conditions are examined using real data.

Discussant:
Mark David Shermis, **Performance Assessment Analytics, LLC**

030. Developing Culturally Relevant Assessment Content: Lessons Learned and the Road Ahead

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

Evolving ideas about fairness in educational measurement have led to greater scrutiny of both construct definitions and test development practices. Whereas legacy approaches focusing on decontextualization may introduce bias favoring the dominant culture, culturally relevant assessments can be developed to affirm cultures and identities, reflect students' lived experiences, disabuse stereotypes, support learning about cultures, and promote social justice. Such efforts, however, are particularly challenging in large-scale assessment programs where standardization (rather than individualization) is the prevailing testing paradigm. This coordinated session includes five presentations from a diverse group of measurement professors and professionals, many of whom work directly with states and local communities. The presentations traverse the past, present, and future of culturally relevant content in large-scale achievement testing programs. This includes reviewing prior research to guide future development of culturally relevant content, illustrating current content development efforts, sharing research on student reactions to culturally relevant content, presenting research methods and results that support cultural validity, and identifying priorities for future research.

Session Organizer:
Jeffrey Steedle, **ACT, Inc.**

Chair:
Joseph A. Rios, **University of Minnesota**

Participants:
Cultural Validity in Large-Scale Assessment: Development and Psychometric Modeling Approaches
Guillermo Solano-Flores, Stanford University
Cultural validity is supported when test development accounts for interactions between students' sociocultural contexts and ways of making sense of and solving items. This presentation covers six practices supporting cultural validity: optimal student sampling, iterative development, cognitive interviewing, analyzing error variance differences across populations, generalizability studies, and disaggregated psychometric analyses.
Developing Culturally Relevant Math and Science Items: Lessons Learned and Student Reactions
Jeffrey Steedle, ACT, Inc.; Cristina Anguiano-Carrasco
This presentation provides lessons learned from ACT's efforts to develop culturally relevant math and science content in the constrained context of high-stakes, admissions testing. This is followed by a summary of reactions to the culturally relevant content gathered through focus group interviews with diverse panels of high school students.
A Case for Community-Relevant Assessments
Pohai Kukea Shultz, University of Hawaii at Manoa; Kerry Englert, Seneca Consulting, LLC
A foundation of cultural and community validity is absolutely necessary if the educational measurement community is truly aiming to develop culturally, linguistically, and community relevant large-scale assessments. This presentation focuses on cultural and community validity evidence gathered for the Kaiapuni Assessment of Educational Outcomes (KĀ'EO).
The Evolution of Inclusive Assessments Through a Multi-Year Research Agenda
Melondy Knight, Curriculum Associates; Kristen Huff, Curriculum Associates
This presentation provides an overview of Curriculum Associates' multi-year research agenda and details how the research can inform inclusive assessment processes and designs. This is followed by a summary of results from initial focus groups and examples of what has been done to attend to student culture in assessment practice.

Discussant:
Kyndra Middleton, **Howard University**

031. Quality Implications of Assessment Engineering in Developing Digital Applications of Testing and Learning

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

Modern psychometricians and assessment service providers increasingly find themselves facing a new paradigm, Assessment Engineering (AE) in which assessment production and management processes can no longer be analogous/manual, but instead can be digitalized/automated by Knowledge Engineering technologies. Although nearly 50 years have passed since the introduction of various AE methodologies (e.g., Evidence-centered Design [ECD], Automatic Item Generation [AIG], etc.) and studies that present a rosy future of AE. Practical applications/implementations of these methods are still expensive, slow, and narrow. One of the reasons is that sufficiently detailed strategies and methodologies have not yet been developed/introduced to effectively/effectively improve the quality (e.g., validity and reliability) of assessment within the AE framework. The concepts and procedures of assessment quality (e.g., validity, validation, or reliability) in traditional assessment development are extended in AE, and the roles/usages of digital technologies are further extended in various AE development targets and processes (e.g., developing procedure for enhancing assessment test security, content validation, reliability assessment, valid formative assessment design). The topics of this session are intended to present several innovative strategies for practical implementations of AE centered around the quality issues (i.e., validity and reliability).

Session Organizer:
Jaehwa Choi, **George Washington University**

Participants:
Test Security Implications of Cloud Collaboration within Assessment Engineering Framework
Seo Young Lee, Prometric LLC; Eunji Lee, George Washington University
Content Validation Implications of Cloud Collaboration within Assessment Engineering Framework
Youn-Jeng Choi, Ewha Womans University; Eunji Lee, George Washington University; Yejin Woo, Ewha Womans University; Hunwon Choi, Ewha Womans University; Yelin Gwak; Sugyung Goh, Ewha Womans University
Development of Adaptive Learning Model within Assessment Engineering: For Valid Formative Assessment
Dayeon Lee
Multivariate Generalizability Theory for Reliability with Item Models: Industrial Mathematics Test Example
Sungyeun Kim, Incheon National University; Sung Kun Yeum, Geunhwa Girls High School; Jinmin Chung, University of Iowa

032. The Lingering Impact of the Pandemic from Multiple Analytic Perspectives

Coordinated Paper Session
10:30 to 12:00 pm
Virtual: Room 4

This coordinated paper session presents an extension of research regarding the impact of the pandemic on key features of an assessment system was evaluated. The extension is designed to highlight some of the lasting impacts of the pandemic over multiple administrations and includes the ongoing application of multiple analytic perspectives. The first paper looks to evaluate the impact of the pandemic on the measurement invariance of an assessment over three years (2019, 2021 and 2022). The second paper concentrates on the degree to which the impact of the pandemic translates to model/data misfit when IRT models are used and whether this impact has changed over multiple administrations. While the third paper is focused on evaluating the impact of the pandemic over the past two years on student performance, the fourth paper looks specifically at the lingering impact of the pandemic at the school level using hierarchical linear models and propensity score matching. This coordinated paper session will look at the lasting impact of the pandemic from different perspectives to provide context for the upcoming implementation of large-scale assessments in 2023 and beyond.

Session Organizer:
Marc W Julian, **DRC**

Chair:
Marc W Julian, **DRC**

Participants:
A Cross-Year Perspective on Measurement Invariance
Huan Wang, Data Recognition Corporation
The Impact of the Pandemic on IRT Model/Data Fit
Christie Plackner, Data Recognition Corporation; Dong-In Kim, Data Recognition Corporation
Assessing Student Performance and Academic Recovery using Propensity Score Matching
Kim Hudson, Data Recognition Corporation; Joanna Tomkowicz, Data Recognition Corporation; Wen-Ching Li, Data Recognition Corporation
The Pandemic Impact on School Performance Using Two Methods
Dong-In Kim, Data Recognition Corporation; Marc W Julian, DRC; Aurore Phenow, Data Recognition Corporation

Discussant:
Karla Egan, **EdMetric LLC**

033. [NCME Book Series] Challenges and Opportunities in Score Reporting

Organized Discussion
10:30 to 12:00 pm
Virtual: Room 5

It has been 3 years since the publication of Score Reporting Research and Applications (https://doi.org/10.4324/9781351136501). This book, part of the NCME book series, includes work in areas such as validity in score reporting, evaluation of subscores, designing and evaluating of score reports for teachers and parents, communicating growth, exploring cognitive affordances of graphical representations, and evaluating the use of interactive reports and dashboards in formative contexts. In this session, we discuss new challenges and opportunities in the area of score reporting that respond to new trends in assessment due to changes in society and education. We see how the field is moving towards reporting systems that can provide teachers and learner with relevant insights based on an abundance of process and response data.

Session Organizer:
Diego Zapata-Rivera, **Educational Testing Service**

Presenters:
Priya Kannan, **WestEd**
April Zenisky, **University of Massachusetts Amherst**
Sandip Sinharay, **Educational Testing Service**
Gavin T. L. Brown, **The University of Auckland**
Linda Corrin, **Deakin University**

034. Investigations to Inform Item Pools and Test Design

Paper Session
1:00 to 2:30 pm
Virtual: Room 1

Chair:
*Hyeri Hong, C*alifornia State University, Fresno

Participants:

Exploring the Effects of Text-Preprocessing Methods in Enemy Item Detection
Yi-Chen Chiang, NABP; Michael R Peabody, National Association of Boards of Pharmacy
The inclusion of enemy items may threaten validity arguments. The study investigates the effects of text-preprocessing methods in enemy item detection. Real data from a licensure examination are used to conduct the analysis and the cosine similarity index was used to measure the similarity between item pairs.

Understanding Factors for Creating Isomorphic Instances in Automatic Item Generation
Danqi Zhu, Fordham University; Yanyan Fu, Graduate Management Admission Council; Kyung (Chris) Han, Graduate Management Admission Council
This study explored factors for creating isomorphic instances in automatic item generation. Using empirical data with 164 items from 35 templates, we found three surface-level features did not yield significant variability of the generated items – using common key (i.e., C), randomization of the key position, and dependency between manipulated elements.

Right-Censored RT Distribution and Speededness: A Case Study on Adjusting Testing Time
Furong Gao, Human Resources Research Organization
This study examines response time (RT) distribution from a CAT test where speededness is observed. Using a fitted lognormal distribution to the observed RT data, the study proposes a method to derive and adjust the test time limit to eliminate or reduce the speededness.

Investigating the Impact of Item Pool Characteristics on Multistage Test Design with Response Time
Hyun Joo Jung, University of Massachusetts Amherst
This study investigates how the recently suggested multistage test design with response time (MST-RT; Park, 2020) is affected by item pool characteristics, such as pool sizes and item difficulty (2020). We investigate the impact of various item pool properties on the MST-RT measurement accuracy.

Investigation of CAT 95% CI Stopping Rule via the Beta-Binomial Model
Johnny Denbleyker, Kaplan; Shuqin Tao, Cambium Assessment, Inc.
A module-based expected percent correct (EPC) stopping rule for CAT is proposed to enable investigating potential termination decisions for variable-length CAT assessments. A licensure CAT test prep assessment is used to empirically illustrate this stopping rule methodology to the existing 95% CI rule currently employed.

Discussant:
Terry Ackerman, University of Iowa

035. Differential Item Functioning: Sources and Detection

Paper Session
1:00 to 2:30 pm
Virtual: Room 2

Chair:
Gabriel Wallin

Participants:

Identification of Differential Item Functioning Using Machine Learning
Tony A Mangino, University of Kentucky; Holmes Finch, Ball State University; Brian French, Washington State University; Cihan Demir, Washington State University
The understanding of group differences in item functioning is complex, yet important to supporting score use. We compare machine learning differential item functioning (DIF) detection techniques to traditional DIF methods for accuracy and efficiency, including in the presence of multiple groups. Machine learning techniques hold promise over traditional methods.

Interaction of Group Ability, Size, and Direction of Bias on DIF Detection
James Weese, University of Arkansas; Ronna Turner, University of Arkansas
DIF simulations are generally designed with larger subgroups being favored and having higher ability. We tested whether DIF detection rates were impacted by group ability, sample size, and direction of item bias using SIBTEST. Significant interactions indicate consistently lower DIF detection in 3PL data when smaller groups have higher ability.

What's the DIF? Item Properties Associated with DIF on the ACT
Jeffrey Steedle, ACT, Inc.; Shalini Kapoor, ACT, Inc.; Shichao Wang, ACT, Inc.
This study used machine learning to identify content, psychometric, and item context variables that were important predictors of DIF on the ACT test. Findings highlight item types that contribute most to achievement gaps and item selection approaches with potential to minimize construct-irrelevant factors contributing to differences in achievement.

Discussant:
Qiwei He, Educational Testing Service

036. Beyond Basketball and Bodegas: Pursuing True Cultural Validity in Formative Assessment

Organized Discussion
1:00 to 2:30 pm
Virtual: Room 3

Tropes related to basketball, clothing, hair, and superfluous community contextual details permeate the narratives that constitute "multicultural" test items from teacher-derived formative assessment to large-scale assessment tools. As instrument developers pursue inclusion and representation through test content, they walk a fine line between being complicit in using racist, ableist, and gendered language and creating a test environment that honors the linguistic and cultural heritage of its intended users. Through the eyes of a community of developers brought together for a new inclusive, equity-informed R&D initiative, this session will feature lessons learned from the implementation of a culture-forward approach to validity for K-12 formative assessment. This panel will highlight how cultural validity must expand to include the positive experiences of multi-generational American students who sit at the intersections of oppressed identities, and why our evaluative processes that govern some of the most critical gateways within their educational experiences must change. Through a dialogic process with attendees, this session will feature how we can collectively tackle the better psycholinguistic approaches in technology-enhanced and technology free assessment prototypes and why community must be a partner in this process in order to expand a new approach to asset-based, culturally representative assessment content. **Session Organizer:**
Temple S Lovelace, **Advanced Education Research and Development Fund**

Presenters:
Lauren Kendall Brooks, **Advanced Education Research and Development Fund**
Karina Rodriguez, **Highlander Institute**
Teaira McMurtry, **University of Alabama, Birmingham**

037. Linking and Equating: Models and Tradeoffs between Sample and Precision

Paper Session
1:00 to 2:30 pm
Virtual: Room 5

Chair:
Hongyu Diao, **Educational Testing Service (ETS)**

Participants:

Linking with the Bayesian Item Response Theory Model
Brandon LeBeau, University of Iowa; Xiaoting Zhong, University of Iowa
Linking is an important consideration when comparing scores over time or across groups. This research will explore the extent to which the Bayesian IRT model can properly link scores across time or groups. This step would estimate item parameters, linking constants, and person parameters in a unified modeling framework that can be flexible to account for other design or explanatory factors.

An Investigation of Small Sample IRT Rasch Model Concurrent Calibrations
Tzu-Chun Kuo, Kaplan North America
Small sample equating has gained increased attention in recent studies due to only low volume of examinees being available. This study compared two maximum likelihood and two fully Bayesian algorithms for estimating small sample Rasch model concurrent calibrations. Varying sample sizes, differences in item difficulties and person abilities were considered.

Sample Size and Estimation Precision When Utilizing the Masters' Partial Credit Model
Michael Custer, Riverside Insights; JongPil Kim, Riverside Insights
Practitioners are accustomed to the trade-off that exists between the costs of obtaining a sample and estimation precision. This study utilizes an analysis of diminishing returns to examine the relationship between sample size and item parameter estimation precision when utilizing the Masters' Partial Credit Model followed by sample size recommendation.

Evaluating Rasch Equating Methods when Sample Size Fluctuates
Isaac Li
Certification exams see rise and fall in examinees per administration, worsened by unforeseen factors like the pandemic. Uneven samples challenge the quality of equating, particularly for low- and medium-sized programs. This study aims to evaluate such methods as the mean/sigma, concurrent, anchor, and TCC under the Rasch framework.

Impact of Item Parameter Drift on Ability Estimates in Computerized Adaptive Testing
Daniel Edi, Pearson Assessments and Qualifications; Sonya Powers, Edmentum, Inc.
This study simulated various levels of IPD and evaluated IPD in an operational CAT with low item exposure. Results indicated that in both the operational CAT and simulated CAT conditions, IPD did not result in meaningful differences in ability and performance level classification accuracy.

Discussant:
Ruben Castaneda, **College Board**

038. Leverage the Partially Confirmatory Approach to Psychometric Modeling with Bayesian Regularization

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

A partially confirmatory approach to psychometric modeling with Bayesian regularization was introduced recently. In this session, four papers are presented to exploit the boundaries of the approach. The first paper investigates the regularized latent variable model framework with structural component, where one can regularize different parameter matrices separately or jointly. It can lead to various research designs depending on the combinations of regularizations. The second paper investigates the partially confirmatory cognitive diagnosis model framework with Bayesian Lasso, where the Q-matrix can be partially specified by the experts and inferred from the response data. Under the framework, the fully expert-defined and data-driven methods of Q-matrix construction can be perceived as two extremes of a continuum with different amount of partial knowledge. The third paper extends the approach to address various bifactor models. It covers both the standard- and extended-types of bifactor models under the exploratory or confirmatory senses, with a partially regularized loading matrix and inherited scalability of the approach. The last paper investigates if we can improve the recommender system with the approach. When incorporating partial knowledge of latent factors, the approach can provide interpretation to help reveal the black box of the learning process in the system.

Session Organizer:
Jinsong Chen, **The University of Hong Kong**

Participants:

Introducing the Regularized Latent Variable Modeling Framework with Bayesian Lasso
Jinsong Chen, The University of Hong Kong

This paper investigates the regularized latent variable model framework with structural component. RLVM extends the multiple-indicator multiple-cause model with Bayesian regularization and local dependence. One can regularize three different parameter matrices separately or jointly and fully or partially. It can lead to various research designs that can be used for different purposes, depending on the combinations of different regularizations.

Q-Matrix Inference in Partially Confirmatory Cognitive Diagnosis Modeling with Bayesian Lasso
Yi Jin; Jinsong Chen, The University of Hong Kong

This paper investigates the partially confirmatory cognitive diagnosis model with Bayesian Lasso, where the Q-matrix can be partially specified by the experts and inferred from the data. The fully expert-defined and data-driven approaches of Q-matrix construction can be perceived as two extremes of a continuum with different partial knowledge.

Accommodating Various Bifactor Models Within the Partially Confirmatory Factor Analysis Framework
Yifan Zhang; Jinsong Chen, The University of Hong Kong

This paper extends the partially confirmatory factor analysis to accommodate various bifactor models under the exploratory or confirmatory senses. For the standard-type bifactor models, the loading matrix can be partially regularized. For the extended-type bifactor models with multiple general factors, two loading matrices can be partially regularized, separately or simultaneously.

Improving Recommender System with the Partially Confirmatory Approach and Psychological Factors
Jinsong Chen, The University of Hong Kong; Yifan Zhang; Zhimin Zou, Wenzhou University

This paper investigates if we can improve recommender system with the partially confirmatory approach. The proposed approach comparable to conventional machine learning methods for recommendation. When incorporating partial knowledge, the approach can provide interpretation to the recommendation, thus help make the black box of the learning process transparent.

Discussant:
Lihua Yao, **Northwestern University**

039. Foundational Competencies in Educational Measurement: How Do Measurement Careers Require Foundational Competencies?

Organized Discussion
3:00 to 4:30 pm
Virtual: Room 2

What are "foundational competencies in educational measurement"? How do educational measurement careers require and further develop these competencies? And how can educational measurement programs support these competencies? In October of 2021, NCME President Derek Briggs charged a 12-member Task Force to "develop and maintain foundational competencies in educational measurement." A year later, the Task Force engaged NCME membership in discussion of a draft report presenting three competency domains and five subdomains, as well as examples of how educational measurement careers and curricula develop these competencies. In this symposium, Task Force members will present their final report on "Foundational Competencies in Educational Measurement," with a focus on how careers and curricula require these competencies. Three discussants who were not members of the Task Force will provide commentary on this report: 1) How do the Task Force's foundational competencies support careers like theirs? 2) Are there foundational competencies necessary for their careers that the Task Force overlooked? 3) How can measurement programs better support the development of the foundational competencies their careers require? This symposium debates how foundational competencies develop in measurement programs and manifest in careers. A complementary, subsequent symposium debates the Task Force's proposed foundational competencies on conceptual and theoretical grounds.

Session Organizers:
Derek Christian Briggs, **University of Colorado Boulder**
Andrew Ho, **Harvard Graduate School of Education**

Presenters:
Terry Ackerman, **University of Iowa**
Howard Everson, **CUNY Graduate Center**
Susan Lottridge, **Cambium Assessment**
Sandip Sinharay, **Educational Testing Service**
Alina A von Davier, **Duolingo**

Discussants:
Debbie Durrence, **Gwinnett County Public Schools**
Leslie Keng, **Center for Assessment**
Michael E. Walker, **Educational Testing Service**

040. An Introduction to Bayesian Statistics

Training Session
8:00 to 12:00 pm
Marriott: Floor 5th - Chicago Ballroom A

Understanding Bayesian statistics components and principles is an important skill for researchers and practitioners of educational measurement. This four-hour workshop presents the basic concepts of Bayesian statistics. Multiple examples will be used to assist understanding the four steps of Bayesian analysis: 1) specifying a prior distribution, 2) summarizing evidence about parameter values using a likelihood function, 3) combining the prior and likelihood to form a posterior distribution, and 4) making inferences. Material will be applied in nature with illustrative examples completed using the MCMC procedure in SAS, but with a focus on principles that can be applied in different software. The intended audience includes practitioners, researchers, consumers of Bayesian analysis, and graduate students studying measurement. Comfort with SAS base programing and procedures will be helpful but not necessary as less than one-fifth of content discussed will use SAS. The presentation format will include a mix of illustrations, discussion, and hands-on examples. As a result of participating in the workshop, attendees will be able to: 1) Articulate the major considerations of a Bayesian analysis, 2) Contrast Bayesian analysis with the Frequentist paradigm, 3) Identify the key components to Bayesian research, and 4) Extend shown examples to more complex models and scenarios.

Presenter:
Brian C Leventhal, **James Madison University**

041. Addressing the Data Challenges of Next-generation Assessments: Data Science Upskilling for Psychometricians

Training Session
8:00 to 5:00 pm
Marriott: Floor 5th - Chicago Ballroom B

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

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

042. An Introduction to Creating Video Games for Measurement: From Design to Analysis

Training Session
1:00 to 5:00 pm
Marriott: Floor 5th - Chicago Ballroom C

Participants will learn about considerations integral to the creation of videogames for measuring player learning, including the affordances of different game mechanics and design choices on gameplay data and how to derive meaningful indicators from gameplay data. We will use a variety of games to demonstrate how particular game mechanics impact the collection of gameplay data, the analyses that consequently can be performed with those data, and what they can reveal about player learning. This introductory session is for people interested in learning more about designing or using games for measurement purposes. It will not cover advanced statistical modeling or data mining. The training session will have three parts. Part 1: Identifying Game Mechanics for Measurement will offer an overview of the relationship between game design and gameplay data. Part II: Extracting Meaningful Events and Indicators from Gameplay Data will offer hands-on experience with the critical analytical process involved in identifying important events and deriving indicators. Part III: Examples of Indicators and Analyses of Gameplay Data will focus on basic data analysis approaches that can be used to make sense of gameplay data. Participants should bring a laptop, tablet, or smartphone to access games for the hands-on activities.

Presenters:
Elizabeth Redman, **UCLA CRESST**
Gregory Chung
Tianying Feng, **UCLA**
Kilchan Choi, **CRESST/UCLA**
Jeremy Roberts, **PBS KIDS Digital**

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

Training Session
8:00 to 5:00 pm
Marriott: Floor 5th - Chicago Ballroom F

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

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

044. Demystify Amazon Web Services (AWS): Cloud Computing, and Psychometric Applications

Training Session
8:00 to 12:00 pm
Marriott: Floor 5th - Chicago Ballroom G

Cloud computing has become increasingly popular over the past few years, allowing people to store a massive volume of data, access the newest version of software and use the virtual machine with unbeatable computing power. 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 enable innovative communication of test results. Given the limitations of existing training materials, this workshop targets attendees who do not come from an IT background. In this workshop, we will cover several AWS core services which can be used to accomplish psychometric analyses, store summary statistics, and display results on a dashboard. Participants do not need to have AWS experience. Upon completion, they will be able to streamline typical psychometric tasks on cloud and communicate findings effectively with other stakeholders inside or outside of their organization. This is a heavy hands-on training and participants are strongly encouraged to create a free AWS account ahead of time and bring their laptops to follow along in order to optimize the learning outcomes from this four-hour training.

Presenters:
Huijuan Meng, **Amazon Web Services (AWS)**
Vinita Talreja, **AWS**
Ye Ma, **AWS**

045. Professional Training for Graduate Students in Measurement

Training Session
8:00 to 12:00 pm
Marriott: Floor 5th - Chicago Ballroom H

This training session will address practical topics for graduate students in measurement to find a job and start a career. First, what to do now while they are still in school to best prepare for a job, which includes the types of training employers look for and how to obtain it (classes, workshops, online training, etc.), how to find a topic and complete a dissertation, how to maximize experiences with networking, internships, social media, and volunteering. Second, how to locate and interview for a job, which includes finding open positions and the application process, including tailoring cover letters, references, and resumes. Third, what to expect in the interview process, including online and in-person interviews, job talks, questions to ask, and negotiating an offer. Last, starting a career, adjusting to the work environment, evaluations, and people (clients, students, co-workers, bosses, mentors, etc.), establishing a career path, work-life balance, dealing with a bad fit, and staying current. The session addresses working-from-home, publishing, professional associations and service, layoffs, and the wide variety of jobs available. The session is interactive, geared to addressing the attendees' particular interests during the session, and providing resource material on all topics as a takeaway.

Presenters:
Deborah J Harris, **University of Iowa**
Nathan Wall, **eMetric**
Yi-Fang Wu, **Cambium Assessment, Inc.**

046. Embedded Standard Setting in Practice

Training Session
1:00 to 5:00 pm
Marriott: Floor 5th - Chicago Ballroom A

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

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

047. Visualizations and Interactive Graphics using R

Training Session
1:00 to 5:00 pm
Marriott: Floor 5th - Chicago Ballroom G

The past decade has seen a vast increase in the types of visualizations used in the media, classroom, articles, and reports. It has also seen an explosion in the use of interactive graphics and dashboards. The free statistical package R and its various packages provide a wide variety of tools for producing them. With you working along through each step on your own laptop, this training session will cover some foundational tools for producing static and interactive graphics in R. The package ggplot2 will be introduced for static graphics, with plotly, ggiraph, and other packages demonstrated for adding interactivity. The course will include a wide variety of examples chosen especially for educational statistics and measurement and will end with an opportunity to work with your own data sets. This course is designed for those who have completed a two-course sequence in quantitative methods. A brief (2-hour) interactive video tutorial in R is provided to participants who have no previous experience in R to be completed before the training session. Participants must bring their own laptop computer; all required software will be provided in advance.

Presenters:
Haley Jeppson, **National Institute of Statistical Sciences**
Brian Habing, **National Institute of Statistical Sciences**

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

Training Session
1:00 to 5:00 pm
Marriott: Floor 5th - Chicago Ballroom H

This training session introduces fundamental knowledge in sequence-based mining methods that could be used to tame complex process data in sequential format associated with timestamps and highlights advanced applications of sequence mining in analyzing process data to better support group-level (in)variance explorations of behavioral patterns in large-scale assessments. Specifically, the Presenters will focus on four subtopics, including (1) how to extract and select gram-based features from clickstream sequence, (2) how to compute sequence distance to identify pairwise sequence similarity, (3) how to integrate timing information into sequence-based analysis, and (4) how to use latent sequence models (e.g., hidden Markov model) to identify latent process states and transitions. During the half-day workshop, participants will be provided with an overview of process data collected from computer-based large-scale assessments, learn about various approaches to analyzing process data with sequence mining methods, and obtain hands-on experience with sequential process data analysis through examples and exercises. Intended audience are researchers, students, and practitioners with basic knowledge of process data and familiarity with R/RStudio/Python and interested in learning or applying data-driven methods to process data analysis.

Presenters:
Qiwei He, **Educational Testing Service**
Esther Ulitzsch, **Leibniz Institute for Science and Mathematics Education**
Bernard Veldkamp, **University of Twente**

049. Board Meeting

4:00 to 7:00 pm
Marriott: Floor 2nd - Old Town

050. Implementing More Student-Centric Measurement Processes: Adventures in Developing the Digital SAT

Coordinated Paper Session
8:00 to 9:30 am
Marriott: Floor 5th - Chicago Ballroom B/C

The SAT is transitioning from a paper and pencil to digital exam. This session will share key assessment design decisions, changes to design processes, and methods for gathering validity evidence in the development of the digital SAT that have been implemented in ways that prioritize the student voice and experience. We will share some challenges as well as innovations and discuss opportunities to make other student-focused improvements going forward. The presentations will focus on more authentic and relevant ways for selecting exam content and context for items, practical changes to how we embed fairness considerations in all aspects of the design and development process, the greater use of cognitive labs and student surveys and focus groups to collect validity evidence and inform test design and item development, and prioritizing real-world predictive validity evidence prior to launch of the exam. The session will also include two measurement experts as discussants, one who works largely on fairness issues and one who works largely on state/K–12 assessments, to share how these student-centered design and process changes connect to desired progress in the measurement field and also where we can look to grow and improve. There will be time for audience questions and feedback.

Session Organizer:
Emily Shaw, **College Board**

Participants:
Engaging Students in the Test Fairness Process
Sherral Miller, College Board; Jay Happel, College Board
Redefining "SAT Words"
Garrett Ziegler, College Board
Using Cognitive Labs to Explore Student Test-Taking Approaches
Jim Patterson, College Board
Gathering Predictive Validity Evidence for Various Audiences
Emily Shaw, College Board

Discussants:
Erika Landl, **Center for Assessment**
Maria Elena **Oliveri, Buros Center for Testing-UNL**

051. Empowering Process Data for Data-Informed Decision-Making in Measurement

Coordinated Paper Session
8:00 to 9:30 am
Marriott: Floor 5th - Chicago Ballroom D

This coordinated session highlights four novel studies in exploring how to empower process data to better inform decision-making in measurement. Our session covers a broad range of process data and showcases both psychometric modeling and data-driven approaches to leverage their potentials. The first paper presents a model-based approach to identify and handle rapid item omissions, which facilitates differentiating the causes of omissions, considers the uncertainty in omission classification, supports investigating person and item characteristics associated with different omission behaviors, and provides flexibility in handling different types of omissions. The second paper introduces a recommender system approach to select test items adaptively based on item scores, sequential process data, and examinees' background characteristics. The third paper demonstrates how to extract information from process data with multidimensional scaling and n-grams that serves as a potential mediator between the group membership and final score and apply exploratory mediation analysis to identify a subset of relevant mediators. The fourth paper evaluates the prediction power of aggregate-level process variables to classify respondents' proficiency levels in problem-solving tasks with machine-learning algorithms. These studies lead a pressing direction of integrating process data in real measurement practice and provide constructive suggestions to enhance measurement accuracy in general.

Session Organizers:
Qiwei He, **Educational Testing Service**
Okan Bulut, **University of Alberta**

Chairs:
Qiwei He, **Educational Testing Service**
Okan Bulut, **University of Alberta**

Participants:

A Model-Based Approach to the Disentanglement and Differential Treatment of Engaged and Disengaged Omissions
Esther Ulitzsch, Leibniz Institute for Science and Mathematics Education; Susu Zhang, University of Illinois at Urbana-Champaign; Steffi Pohl, Freie Universitat Berlin
Adaptive Item Recommendation Using Process Data and Examinee Background Characteristics
Okan Bulut, University of Alberta; Seyma N. Yildirim-Erbasli, Concordia University of Edmonton; Surina He; Bin Tan, University of Alberta; Yizhu Gao
Explaining Performance Gaps with Problem-Solving Process Data via Exploratory Mediation Analysis
Susu Zhang, University of Illinois at Urbana-Champaign; Xin Wei, SRI International
Evaluating Prediction Power of Process Variables on Problem-Solving Proficiency Levels with Machine Learning Methods
Qiwei He, Educational Testing Service; Qingzhou Shi, University of Alabama; Francesca Borgonovi, University College London; Organisation for Economic Co-operation and Development; Marco Paccagnella, OECD

Discussant:
Samuel Greiff, **University of Luxembourg**

052. Improving Assessment Decisions Using Collateral Information About Incorrect Responses and Response Times

Coordinated Paper Session
8:00 to 9:30 am
Marriott: Floor 5th - Chicago Ballroom E

These four conceptually related papers describe various models for capitalizing on two types of collateral information in item responses: incorrect responses and response times. They report on evaluations of the models and/or the collateral information using real and simulated data. The first paper describes a class of cognitive diagnostic models that use both correct and incorrect responses to model how students assemble cognitive pieces (called facets) to reach an understanding of a target domain. The second paper describes a reading assessment with one dimension to describe overall reading comprehension and a second dimension based on incorrect responses to produce a diagnostic classification for struggling readers. The third paper reports a study comparing models for incorporating response times with self-reports of functional cognitive behavior. A model with dichotomized response times improved estimation of both item and person parameters. Whereas the first three papers envision tests that are diagnostic for people who are, in some sense struggling, the last envisions using response time information to identify otherwise good readers who may still need to improve the automaticity of their reading for the purpose of reading-to-learn. It examines the validity of collateral response time information in predicting subsequent performance on a statewide exam.

Session Organizer:
Mark Davison, University of Minnesota

Participants:

A New Facet Model for Extracting Instructionally Useful Information from Diagnostic Assessment
Chun Wang, University of Washington
This paper describes a class of cognitive diagnostic models that use both correct and incorrect responses to model how students assemble cognitive pieces (called facets) to reach an understanding of a target domain. Feasibility and estimation are demonstrated with real data from a physics unit on forces and motion.

A Diagnostic Model for Adaptive Assessment and Diagnosis of Complex Cognitive Processes
Joseph DeWeese, University of Minnesota-Twin Cities; David Weiss, University of Minnesota; Ozge Ersan, University of Minnesota-Twin Cities; Mark Davison, University of Minnesota; Patrick Kennedy, University of Oregon; Gina Biancarosa, University of Oregon
This paper describes a reading assessment with one dimension to describe overall reading comprehension and a second dimension based on incorrect responses to produce a diagnostic classification for struggling readers. Simulation and real data are presented on measurement precision and classification accuracy.

Incorporating Response Times into Multidimensional Rating Scale Measurement
Shiyang Su, University of Central Florida; Chun Wang, University of Washington; David Weiss, University of Minnesota
This paper describes assessments that combine polytomous responses and response times in reports of applied cognition, daily activity, and mobility. It compares models that use (a) response times scored continuously, (b) response times scored dichotomously, and (c) no response times. It supports the usefulness of dichotomous RTs for improving precision.

Predicting Reading Proficiency with Response Times on an Online Multiple-Choice Comprehension Assessment
Yun Leng Wong, University of Minnesota; Mark Davison, University of Minnesota
This paper envisions using response time information to identify otherwise good readers who may still need to improve the automaticity of their reading for the purpose of reading-to-learn. It examines the validity of collateral response time information in predicting subsequent performance on a statewide exam.

Discussant:
Stephen G Sireci, **University of Massachusetts, Amherst**

053. Assessing Collaborative Problem Solving at Scale: Individual Contribution to Teamwork

Coordinated Paper Session
8:00 to 9:30 am
Marriott: Floor 5th - Chicago Ballroom F

As one of the core 21st century skills, collaborative problem solving (CPS) is crucial for success in both academia and workplace. However, assessment of CPS is very challenging due to the nature of the constructs, and there is a lack of generally available assessment instruments for CPS. In this coordinated session, we collected five presentations to introduce ETS' recent efforts towards a scalable assessment of CPS. We show how we prepare for assessing CPS from technology, data and psychometrics; how to design the study to measure individual contribution to teamwork; how to address the logistic challenges for data collection and human annotation; and how to leverage AI technology to transcribe audios and identify speakers.

Session Organizer:
Jiangang Hao, **Educational Testing Service**

Participants:

Assessing CPS at Scale – EPCAL Ecosystem and Psychometric Considerations
Jiangang Hao, Educational Testing Service; Emily Kerzabi, Educational Testing Service; Patrick Charles Kyllonen, ETS
Assessing Individual Contribution to Teamwork: Design and Findings
Patrick Charles Kyllonen, ETS; Jiangang Hao, Educational Testing Service; Emily Kerzabi, Educational Testing Service; Yuan Wang, ETS; Rene Lawless, ETS **Large-scale Data Collection of Collaborative Tasks: Challenges and Strategies**
Yuan Wang, ETS; Emily Kerzabi, Educational Testing Service; Rene Lawless, ETS
Developing and Implementing Human Coding of Collaborative Communication at Scale
Emily Kerzabi, Educational Testing Service; Rene Lawless, ETS; Yuan Wang, ETS
Speaker Detection and Automated Transcription of Audios in Collaborative Tasks
Michael Fauss, ETS; Jiangang Hao, Educational Testing Service

Discussant:
Alina A von Davier, **Duolingo**

054. Research Blitz: IRT Models

Research Blitz Session
8:00 to 9:30 am
Marriott: Floor 5th - Chicago Ballroom G/H

Chairs:
Jose Felipe Martinez, **UCLA - School of Education and Information Studies**

Participants:

A New Approach for Item Fit Statistics to Overcome Model-Dependency and Sample-Dependency
Jihang Chen, Boston College; Louis Roussos, Cognia; Liuhan Sophie Cai, Cognia

Evaluating the model-data fit of IRT models is important before we draw inferences from them. A new hypothesis testing approach for improving item fit statistics is proposed to overcome previous issues of model-dependency and sample-dependency. Results suggest this approach can accurately identify the items with a poor model-data fit.

Using a Bi-factor Model to Validate a 3rd Grade Formative Science Assessment
Kayla Bartz, Michigan State University; Lydia Bradford, Michigan State University

This paper examines the use a bi-factor model to validate a 3rd grade formative science assessment. Comparing a unidimensional, 3 factors: first order, and a bi-factor model found that the bi-factor model was the best fit. Rather than showing a pass/fail, it allows a deeper assessment of students understanding.

IRT Misspecification and Its Implications in the Modeling of Recognition Task Data
Qi Huang, University of Wisconsin - Madison; Daniel Bolt, University of Wisconsin, Madison

Traditional item response theory (IRT) models, like the 2PL, are often indiscriminately applied to tests with dichotomously scored items without considering the type of underlying response processes. Using a recognition test as example, we demonstrate the potential consequences of 2PL misspecification in relation to a study of differential item functioning.

Using Topic Models to Characterize Mixture IRT Latent Classes
Constanza Mardones, University of Georgia; Allan Cohen, University of Georgia

The purpose of this study was to use the supervised latent Dirichlet allocation (sLDA) to help characterize latent classes obtained by a mixture IRT model. Preliminary results suggested that topics provided by sLDA reflected the different reasoning examinees used for each class.

Comparing Priors for Estimating Sparse Ordinal Indicators in Bayesian Factor Analyses
Sonja D Winter, University of Missouri, Columbus; Jorge Sinval, ISCTE — University Institute of Lisbon; Edgar Merkle

A common issue in educational measurement is low item endorsement of extreme response options. Modeling such sparse data can result in non-convergence, overly optimistic model fit indices, and biased parameter estimates. This study examines the potential of the Dirichlet prior distribution to model such data using Bayesian estimation.

Modeling Not-Reached Responses in the Problem Solving and Inquiry Tasks of TIMSS 2019
Yuan-Ling Liaw

Respondents may not attempt the end of a test due to time limits or lack of motivation. This paper aims at investigating the not-reached responses in the Problem Solving and Inquiry (PSI) Tasks of TIMSS 2019 from task and item characteristics and test-taking behavior across countries.

Modeling Context Characteristics for Contextualized Assessment: A Bayesian Contextualized Item Response Model
Nixi Wang, University of Washington; Min Li, University of Washington; Klint Kanopka; Dongsheng Dong, University of Washington; Philip Hernandez; Maria Araceli Ruiz-Primo, Stanford University

The design of contextualized assessment is commonly used and yet, calibrating item contexts is still lacking in practice. This paper addresses the sociocognitive and sociocultural characteristics of contextualized items in a testlet-based physics assessment for multiple levels of context, and examines corresponding effects on students' ability of solving tasks.

An Overview of Approaches to Violations of Local Item Independence in Testlets
Yan Yan, Georgia Tech; Kirk Becker, Pearson

This work reviewed the researches on testlet-based assessment and summarized potential modeling approaches for testlets. A simulation study was run to compare the performance of several models on testlets under different data generation methods. It showed that in certain conditions, the standard IRT models can still work with testlet data.

055. Moving Towards an Equitable and Just Profession: Lessons Learned from the Field

Organized Discussion
8:00 to 9:30 am
Marriott: Floor 5th - Denver/Houston

Many scholars and advocates have justifiably critiqued the slow pace of the educational measurement field in addressing inequities especially as it relates to racism. After all, Edmund Gordon (1995) and others (e.g., Hood, 1998; Gould, 1981) have been writing about inequities in our field for more than 30 years. Currently, there are a number of efforts underway that aim to make educational assessment more equitable and just. This discussion session highlights efforts that are focusing on diversifying the pool of professionals who work in our field to reflect the population of test takers. More specifically, our session features the efforts of three organizations committed to improving the representation of women and people from marginalized communities in educational measurement. Each of the three organizations—The Center for Measurement Justice, Women in Measurement, and The Center for Assessment—are trying to address different aspects of this work, while navigating multiple challenges and gaining critical perspectives. This "organized discussion" will feature brief, introductory presentations by representatives of the three organizations and a facilitated, transparent discussion about our early successes and key, critical challenges we face as we work to scale our efforts.

Session Organizer:
Scott Marion, **National Center for the Improvement of Educational Assessment**

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

Presenters:
Jade Caines Lee, **University of Kansas**
Maria Hamdani, **Center for Measurement Justice**
Susan Lyons, **Lyons Assessment Consulting**

056. Design and Evaluation of Adaptive Testing in Large-Scale Survey Assessments

Coordinated Paper Session
8:00 to 9:30 am
Marriott: Floor 5th - Los Angeles/Miami

Large-scale survey assessments (LSAs) are important tools for supporting educational policy decisions. Adaptive testing may further enhance such decisions by improving measurement precision and several LSAs now have experience with adaptive testing. This experience can be through special studies (Oranje et al., 2014; Wu, 2017), a full implementation of multistage adaptive testing (MST) as in the case of the Programme for International Student Assessment (PISA; Yamamoto et al., 2019) and the Programme for the International Assessment of Adult Competencies (PIAAC; Yamamoto et al., 2018), or a special form of group-level adaptive testing such as in the Progress in International Reading Literacy Study (PIRLS; Mullis & Martin, 2019) and Trends in Mathematics and Science Study (TIMSS; Mullis et al., 2021). In LSA, the goal is to report group-level proficiencies rather than individual proficiencies through elaborated sampling and testing designs. This focus on group rather than individual scores has an impact on how best to implement and evaluate adaptive testing methodologies. The emphasis in this coordinated session is to provide an up-to-date overview of results and insights obtained with adaptive testing designs across various well-known LSAs.

Session Organizer:
Peter van Rijn, **ETS Global**

Chair:
Peter van Rijn, **ETS Global**

Participants:
Can Adaptive Testing Improve Testing Experience in Educational Survey Assessment?
Yi-Hsuan Lee, Educational Testing Service; Yue Jia, Educational Testing Service
Adaptive Designs in TIMSS and PIRLS - Improving Accuracy in Survey Assessments
Matthias Von Davier, Boston College
Optimal Multistage Adaptive Test Design in PISA: Improving Group-Level Inference
Peter van Rijn, ETS Global; Usama Ali, Educational Testing Service; Hyo Jeong, Sogang University; Frederic Robin, ETS
Stepwise Assembly for Multistage Adaptive Testing Designs in PISA and PIAAC
Usama Ali, Educational Testing Service; Peter van Rijn, ETS Global; Frederic Robin, ETS

Discussant:
Hyo Jeong, **Sogang University**

057. eBoard Session 1

Electronic Board Session
8:00 to 9:30 am
Marriott: Floor 7th - Salon I

Participants:

1. **Can Online Exams Provide Meaningful Assessment of Student Learning?**
Dahwi Ahn, Iowa State University; Jason C.K. Chan, Iowa State University
The prevailing wisdom amongst many instructors is online exams are too easy and undiagnostic. We investigated whether online exams are indeed less diagnostic than in-person exams in a high stake, naturalistic environment. Specifically, we compared exam scores during Spring 2020, in which exams switched from in-person to online administration mid-semester.
2. **Optimal Use of Technology-Enhanced Items in an Interim CAT**
Jinah Choi, Edmentum, Inc.
This study investigates the optimal use of technology-enhanced items in a computerized adaptive test by comparing TE item difficulty to examinee ability. The psychometric properties of TE and MC items are compared, and simulations are used to evaluate the impact of TE items on testing time and ability estimation accuracy.
3. **Exploring Performance-Based Heuristics for Adapting Digital Instruction**
Fusun Sahin, American Institutes for Research; Sebastian Moncaleano, Curriculum Associates, LLC; Logan Rome, Curriculum Associates
This study considers how student interactions within digital instruction can be used to evaluate student performance on reading or math skills without the need for formal assessment at the end of the lesson. Results showcase how process data can be harnessed to build a heuristic algorithm for adapting digital instruction.
4. **The Categorisation and Analysis of Multiplication Errors in a Digital Assessment**
Rosa Leino, Standards & Testing Agency; Liam James Maxwell, Standards & Testing Agency; Fusun Sahin, American Institutes for Research
A categorisation framework was developed to examine the types and prevalence of errors produced by pupils in an on-screen assessment measuring fluent recall of multiplication tables through a set of timed questions. The results indicate that construct-irrelevant keystroke errors comprise a small quantity (0.4%) of the responses.
5. **An Exploration of Test-taking Behaviors of Multilingual Learners in TIMSS 2019**
Jung Yeon Park, George Mason University; Angela Miller, George Mason University; Sujin Kim, George Mason University
This study explores relationships between multilingual learners' test-taking behaviors and their performance on math and science tests to identify unique needs of the minority group for STEM education. Specifically, data from a digital version of TIMSS 2019 were used to examine differential response time functioning between different home language groups.
6. **Comparisons of Field Test Item Estimation Methods on the MCAT® Exam**
Marc Kroopnick, Association of American Medical Colleges; Ying Jin, Association of American Medical Colleges; Bethany Bynum, HumRRO
The study examined several field-test (FT) item parameter estimation methods on the MCAT® exam. The findings suggest that one method, which involves a free, concurrent calibration of FT and operational items, outperforms the others in yielding the FT item parameter estimates closest to their post-equating operational estimates.
7. **Score Linking Using Student Background Variables**
Ruoyi Zhu, University of Washington; Ying Lu, College Board; Amy Hendrickson, College Board
There is sometimes the need to conduct score linking when the commonly used linking designs (e.g., random equivalent group design or common item linking design) cannot be implemented. This paper evaluates and compares two linking methods using examinee background variables: propensity score weighting (PS) and pseudo-equivalent group (PEG) linking.

8. **Examining Two Scaling Methods for Weights in Multilevel Modeling**
   Alexandra Lane Perez, University of Connecticut; Katherine Furgol Castellano, Educational Testing Service; Jonathan Weeks, Educational Testing Service; Matthew Johnson, ETS; Daniel Mccaffrey, Educational Testing Service
   Currently no studies examine the performance of the two recommended scaling methods for level one weights in MLM when weights are for matched samples. Therefore, we conducted a simulation study examining this. Overall, the two scaling methods performed similarly in the two-level MLM, but we will extend to three-levels.
9. **Leveraging Exploratory Multivariate Techniques and Intervention Fidelity Data to Identify Active Ingredients**
   Jay Jeffries, University of Nebraska-Lincoln
   Treatment fidelity is an important concept for identifying under which conditions a treatment is effective, though few utilize its data. This study leverages exploratory multivariate techniques to investigate fidelity measures and reveal program active ingredients. Fidelity measures were systematically evaluated to uncover intervention elements that merit pronounced attention during treatment.
10. **Exploring Students' Utilization of the "Don't Know" Response in Financial Knowledge Questions**
    Katrina Borowiec, Boston College; Angela Boatman, Boston College
    Financial management skills are essential when financing postsecondary education. Using an experimental design (n=842), this study explores whether including a "don't know" response option impacts students' responses to new measures of general financial and financial aid knowledge. Offering the "don't know" option is associated with fewer items answered correctly.
11. **Scale of Critical Practice in Content-Integrated Education for Multilingual Learners: Psychometric Properties**
    Jung Yeon Park, George Mason University; Sujin Kim, George Mason University; Xiaowen Chen, George Mason University; Bilgehan Ayik, George Mason University; Yixin Zan, George Mason University; Woomee Kim, George Mason University; Dai Gu, George Mason University
    This study examines psychometric properties of a scale developed to measure teacher perceptions of their pedagogical practice toward multilingual learners in general content classrooms. Exploratory factor analysis was conducted to identify the underlying factor structure. Measurement invariance between different types of teachers is tested in the framework of EFA.
12. **The Development of a Mathematics Instruction Observation Tool: A Validity Argument**
    Elizabeth R. Thomas, Southern Methodist University; Leanne Ketterlin Geller, Southern Methodist University; Erica Lembke, University of Missouri; Sarah King, University of Texas at Austin
    The Mathematics Instruction Observation Tool (MIOT) was developed to support the on-going coaching of middle school mathematics teachers implementing data-based individualization and evidence-based practices to improve student outcomes. The focus of this presentation will be on the initial development and validation of the tool.
13. **Effect of Using Rubrics on Motivation and Performance: A Meta-Analysis Study**
    Sandra Liliana Camargo Salamanca, Purdue University; Fabio Andres Parra-Martinez, University of Arkansas; Ammi Chang, Purdue University; Yukiko Maeda, Purdue University
    In this meta-analysis, we examined the effect of using rubrics on motivation and performance in K-16 formal learning settings across languages, including reviewing empirical reports in English, Spanish, Portuguese, and Korean. Preliminary results support the use of rubrics as a tool to enhance academic performance and motivation in the classroom.
14. **Data Visualization for Telling Policy Stories on Changing Caribbean High-Stakes 11+ Testing**
    Jerome De Lisle, University of the West Indies; Tracey Michelle Lucas, University of the West Indies; Murella Sambucharan-Mohammed, The University of the West Indies; Carla Kronberg, UWI St Augustine; Sharon Phillip, The University of the West Indies; Nalini Ramsawak-Jodha, University of the West Indies, St. Augustine, Trinidad; Nisha Harry, The University of the West Indies
    We constructed twelve infographics telling stories supporting policy change for 11+ testing in Trinidad and Tobago. We hypothesized that popular narratives and myths inherited from colonial times have maintained the system. Several national databases were used to construct counter-narratives using static infographics. Evaluation data provide insight on design and sensemaking.
15. **A Content Analysis of Documentation and Psychometric Evidence for New Test Editions**
    Miriam Crinion, Buros Center for Testing - University of Nebraska-Lincoln; Jessica L. Jonson, Buros Center for Testing-UNL
    This presentation will share findings from a content analysis that examined the extent to which the technical manuals and independent test reviews provided test users with the documentation and psychometric evidence needed to determine if a new edition of commercially-available cognitive and achievement tests should be adopted.
16. **Methodological Developments in Quantifying the Reliability of Accountability Scores for School Identification**
    Lily An, Harvard Graduate School of Education; Brian Gong, Center for Assessment
    School accountability scores aggregate student level results to support policymakers' decision-making into which schools require additional supports. The sampling error in student test scores has implications for error in school accountability scores. This paper advances methods for estimating the reliability of accountability scores for identification by analyzing error over replications.
17. **Understanding Admissions Websites: Differences in Holistic Language and Requirements**
    Jose de Jesus Sotelo, Educational Testing Service (ETS); Reginald M Gooch, Educational Testing Services; Guangming Ling, Educational Testing Service; Kevin Williams, Educational Testing Service
    Despite the increased popularity of the holistic admissions approach, there is little consensus on its definition and limited knowledge of its practical implementation at institutions. In this study, we analyzed the admissions web pages of 150 U.S. postsecondary institutions to understand holistic admissions policies as communicated to prospective college students.

18. **Sketching a Validity Argument from User Experiences Shared on Twitter**
Sergio Araneda, University of Massachusetts Amherst; Stephen G Sireci, University of Massachusetts, Amherst
This paper presents a way to sketch a validity argument based on user experiences, presenting a case of a test in Chile and using the experiences shared on Twitter to sketch a validity argument as part of an Experiential Approach to Test Validation.

19. **Assessing The Performance of Smoothed-Bootstrapping in DIF Detection with Small Sample Size**
Yongseok Lee, University of Florida; Ziying Li, University of Florida; Matthew Faiello, University of Florida; Anne Corinne Huggins-Manley, University of Florida; Mary Bratsch-Hines, University of Florida
The issue of measuring item fairness using differential item functioning (DIF) techniques in small sample sizes has not yet been resolved. In this study, we evaluate the performance of smoothed bootstrapping with SIBTEST for small sample DIF detection through a Monte-Carlo simulation.

058. Advances in Item Response and Response Time Modeling

Paper Session
9:50 to 11:20 am
Marriott: Floor 5th - Chicago Ballroom B/C

Chair:
Ismail Dilek

Participants:

Evaluation of Item Fit and Item Parameter Drift Using Posterior Expectations
Li Cai, UCLA; YoungKoung Kim, College Board; Yun-Kyung Kim, UCLA
This study proposes to use posterior expectations to evaluate item fit and item parameter drift. It improves on existing methods by taking into account the uncertainty of item parameter estimates with a diminished computational burden. The performance of the method is demonstrated by simulation study and an empirical example.

Impact of Modeling Item Response Times on Mathematical Achievement Test Scores
Susan Embretson, Georgia Institute of Technology
Item response times are increasingly available on achievement tests. Several models have been proposed to incorporate response times with IRT-based accuracy models. The impact of these models on mathematical achievement estimates in middle school was examined.   The results indicate that mixture-based models provided more added information than the other models.

A Quasi-Poisson Testlet Model for Count Data
Cornelis Potgieter, Texas Christian University; Xin Qiao, Southern Methodist University; Akhito Kamata, Southern Methodist University
We propose a semiparametric latent variable model for unbounded count data collected from testlet-based educational assessments. The model accounts for both testlet effects and various dispersion levels in the data. We present the model formulation, a moment-based estimation method, and an illustrative implementation using real data from a large-scale assessment.

Applying the Intermodel Vigorish to Quantify the Value of Item Response Modeling
Ben Domingue, Stanford University; Klint Kanopka; Radhika Kapoor; Steffi Pohl, Freie Universitat Berlin; R. Phillip Chalmers, York University; Charles Rahal, University of Oxford; Mijke Rhemtulla, University of California, Davis
Deployment of item response models necessitates assessment of model fit. We introduce the InterModel Vigorish (IMV) to quantify performance based on improvement in predictive accuracy between two models. The IMV's values are generalizable and can be used to compare non-nested models. We illustrate its utility using simulated and empirical data.

Estimating Testing Time through Bayesian Stochastic Modeling with PyMC and Bean Machine
Yi-Fang Wu, Cambium Assessment, Inc; Zhongtian Lin, Cambium Assessment, Inc
In response to the call for fewer and smarter assessments, the study tackles a practical need—reducing testing time—using data-driven approaches. We adopt response time models by van der Linden and conduct Bayesian stochastic modeling to model identification and estimation with new probabilistic programming environments, PyMC and Bean Machine.

Discussant:
Alexander Weissman, **Law School Admission Council**

060. Latest Work in Item Difficulty Modeling and Cognitive Complexity

Coordinated Paper Session
9:50 to 11:20 am
Marriott: Floor 5th - Chicago Ballroom E

Item difficulty modeling (IDM) studies involve predicting item difficulty and discrimination from item response demand features (e.g., content area knowledge and skill requirements). A goal of this research is to validate item response demands as significant predictors that can be used to (a) engineer items to ensure that they elicit intended, construct relevant cognitive processing (i.e., intended declarative and procedural knowledge); (b) increase the likelihood that observed item difficulty will match targeted test scale locations; (c) guide item specification, item writing, and test forms assembly; and (d) aid automated item generation. IDM methodologies include OLS regression, regression trees, the linear logistic test model and, most recently, machine learning and natural language processing techniques. Currently, no generally accepted definitions of item cognitive complexity exist. A popular hypothesis is the complexity and difficulty are related and separable. In this session, four groups who are active in IDM and cognitive complexity research will present on their latest thinking and work. A discussant will interpret the studies to focus on applications of findings for operational test development and for validity arguments for items included in operational test forms, which will help the session further embody the conference theme, Leveraging Measurement for Better Decisions.

Session Organizer:
Steve Ferrara, **HumRRO**

Participants:

Investigating Reading Comprehension Construct Stability Using IDM
Christina Schneider, Cambium Assessment, Inc.; Jing Chen, Cambium Assessment
We will create and validate item difficulty models using item metadata from a Grade 3–8 reading assessment that include text complexity requirements and determine if those models generalize to items administered in 2022 when the state reduced the text complexity associated with items in each grade.

Reinvigorating Webb’s Depth of Knowledge in Three Content Areas
Marjorie Wine, ATLAS: University of Kansas; Alexander Hoffman, Aledev Consulting
Webb’s (2002) DOK typology of cognitive complexity is widely used, but industry use has drifted from his original intent, losing its central thrust on automaticity vs. deliberation. We reinvigorate that idea across three content areas – CCSS ELA, CCSS mathematics, NGSS science – by recognizing the impact of increased proficiency on cognitive complexity.

An IDM Study of Mathematics Item Predictors Using a Representation and Manipulation Framework
Kathryn Nicole Thompson, James Madison University; Steve Ferrara, HumRRO
We examined whether significant predictors of item difficulty in previous mathematics studies generalize to 8th grade algebra items using a representation and manipulation framework. Item predictors explained a greater amount of variance in item difficulty compared to item discrimination. Implications for test development and item writing will be discussed.

The Influence of Opportunity to Learn in Item Difficulty Modeling Studies
Steve Ferrara, HumRRO; Tony Albano, University of California, Davis; Jeffrey Steedle, ACT, Inc.; Mark Johnson, Cognia, Inc.
Some portion of unexplained variance in predicted item p-values may be explained by differential OTL across schooling levels. The variance explained by OTL could suggest a limit on the variance we can explain using construct relevant item response demands that are used as difficulty predictors in other studies.

Discussant:
Catherine Close, **Renaissance Learning**

061. Cognitive Diagnostic Modeling: Mathematical Issues and Model Specifications

Paper Session
9:50 to 11:20 am
Marriott: Floor 5th - Chicago Ballroom F

Chair:
Benjamin R. Shear, **University of Colorado Boulder**

Participants:

Testing the Invariance of Learning Trajectories in CDM: A Two-Step Estimation Approach
Zechu Feng, The University of Hong Kong; Qianru Liang, The University of Hong Kong; Hulya Duygu Yigit; Jimmy de la Torre, University of Hong Kong
When estimating learning trajectories in CDMs, the invariance of attribute transition probabilities and independence among the attributes cannot always be satisfied, and thus need to be tested, In this study, a two-step likelihood-based procedure to test these assumptions is proposed. Simulation results indicate that the proposed method is promising.

Mathematical Issues Impacting the Fitting of Latent Variable Models
Eric Loken, University of Connecticut; Jeremy Teitelbaum, University of Connecticut
Generically identified models may have certain regions where subsets of the parameters are not identified. Using three examples from factor analysis and latent class analysis, we show that the problematic regions can greatly affect estimation and inference. We can avoid fitting non-identified models, but sometimes we cannot avoid finding them.

An Exploratory Markov Network for Modeling Local Item Dependency in Diagnostic Assessments
Hyeon-Ah Kang, University of Texas at Austin; Jingchen Liu, Columbia University; Zhiliang Ying, Columbia University

The study proposes an exploratory approach to modeling local item dependency in cognitively diagnostic assessments. We integrate a diagnostic classification model with a Markov network such that inter-item dependency can be modeled via an undirected graph. Empirical validation suggests that the proposed approach holds potential as a robust analytical framework.

The Effects of Measurement and Structural Model Misspecifications in Longitudinal Diagnostic Classification Models
Matthew James Madison, University of Georgia; Meghan Fager, Hitachi Solutions America; Allen Christopher Moore, University of Georgia; Selay Zor, University of Georgia

When applying longitudinal DCMs, there are several model specifications that are made based on the design and goals of the assessment. This study uses a simulation study to examine the effects of measurement and structural model misspecifications. Results suggest robustness when one type of misspecification occurs, but not both.

Examining the Effects of Retrofitting Diagnostic Models to Item Response Theory Data
Allen Christopher Moore, University of Georgia; Matthew James Madison, University of Georgia

Retrofitting diagnostic models to non-diagnostic assessments has not been thoroughly explored in psychometrics. Understanding the impacts of retrofitting could better inform the use of diagnostic models when retrofitting is the only option. This study seeks to examine the impacts of retrofitting on item parameters and model and item fit indices.

Discussant:

Mubeshera Tufail

062. Internships in the Measurement Profession: A Discussion Among Organizers, Mentors, and Students

Organized Discussion
9:50 to 11:20 am
Marriott: Floor 5th - Chicago Ballroom G/H

An internship has become an increasingly common experience for measurement graduate students. There is great variety in types on internships, roles for interns, modality of delivery (virtual, in-person, hybrid) and experience for student interns. During this engaging discussion, three professional mentors/organizers of internships and two graduate students who have completed internships will discuss the measurement internship experience. Specifically, they will describe their experience, the benefits for students, tips for earning an internship, and why internships are so valuable to employers and students. The session will be appropriate for graduate students interested in internships, educators to understand the value of internships for students, and professionals looking to start or improve an internship experience at their institution or company.

Session Organizer:
Brian C Leventhal, **James Madison University**

Moderator:
Brian C Leventhal, **James Madison University**

Presenters:
Sarah Alahmadi, **James Madison University**
Joshua Goodman, **NCCPA**
Janine Jackson, **Morgan State University**
Xin Li, **ACT, Inc.**
Christopher Runyon, **NBME**

063. Automatic Generated Items and Automatic Enemy Item Detection

Paper Session
9:50 to 11:20 am
Marriott: Floor 5th - Denver/Houston

Chair:
Kylie Gorney, **University of Wisconsin-Madison**

Participants:

Automatic Item Paraphrasing to Avoid the Repetitive Use of the Same Items
Seyma N. Yildirim-Erbasli, Concordia University of Edmonton; Guher Gorgun, University of Alberta; Okan Bulut, University of Alberta

In this study, we designed a paraphrased question generator to address the issue of retesting with the same test form. We fine-tuned the pre-trained T5 model on the task of paraphrase generation. Our study shows promising results to ask a question with identical meaning but with different words under various test settings.

Enemy Item Detection for Quantitative Item Type
Yanyan Fu, GMAC; Kyung (Chris) T. Han, Graduate Management Admission

Several natural language processing (NLP) based similarity metrics were used to identify the enemy pairs for quantitative item types. The preliminary results showed that enemy pairs had higher similarity than non-enemy pairs, indicating that the similarity metrics could effectively identify enemy pairs for different quantitative item types.

An Exploration of Natural Language Processing for Enemy Item Detection
Liu Liu, University of Washington; Marcus Walker, National Commission on Certification of Physician Assistants; John Weir, National Commission on Certification of Physician Assistants

Enemy items share similar content/characteristics or give clues for answering another item, compromising measurement precision and diminishing test validity. This study explores the effectiveness and evaluation of the performance of Natural Language Processing (NLP) techniques: word embeddings and classification algorithms to identify enemy items in an operational certification item bank.

A Method for Banking Large Numbers of Generated Items
Hollis Lai; Tahereh Firoozi, University of Alberta; Mark J Gierl, University of Alberta

The purpose of our paper is to describe and demonstrate a new method for organizing, accessing, and monitoring generated items using content coding. We provide a conceptual overview of content coding for banking generated items. We also provide a methodology for generating items using content codes.

Exploring Pretesting Designs for Automatic Generated Items
Fen Fan, NBME; Amanda Clauser, National Board of Medical Examiners; Thai Quang Ong, National Board of Medical Examiners

We will explore matrix sampling designs for pretesting AIG items to establish guidelines where the Rasch model is robust to inclusion of AIG item families on a single form (local independence). Testing organizations can leverage results to maximize number of AIG items that can be pretested on a single form.

Discussant:
Lihua Yao, **Northwestern University**

064. Issues and Strategies in Maintaining Testing Programs Internationally and in Various Languages

Coordinated Paper Session
9:50 to 11:20 am
Marriott: Floor 5th - Chicago Ballroom D

There are many technical and operational challenges of maintaining a testing program internationally and in various languages. The Standards are light in providing guidelines in these contexts. This session will provide a set of issues and solutions using real-world examples as a way of offering information to the audience and inform practice. The four presentations will provide the information and examples in various contexts and different testing programs. The first presentation will be in certification across countries involving (a) job analysis, (b) policy issues, (c) test blueprints, (d) item development for use in different languages, and (e) psychometric procedures including DIF. The second presentation will talk about the alignment of an international assessment program across 100 countries presenting the methodology and results. The third presentation will present the methods and results of maintaining new forms in two languages across three countries while ensuring comparability of content and scale. The fourth presentation will provide the technical and policy issues and solutions for implementing testing programs internationally online. Finally, a discussant will put these presentations into a perspective and provide an overview of what testing programs should do when testing internationally and/or in different languages.

Session Organizer:
Thanos Patelis, **Johns Hopkins U & University of Kansas**

Chair:
Thanos Patelis, **Johns Hopkins U & University of Kansas**

Participants:

Challenges and Real-World Examples of Addressing Issues when Testing Internationally
Andrew Wiley, ACS Ventures; Chad W. Buckendahl, ACS Ventures, LLC

International Standards Alignment Research Considerations
Jennifer Merriman, International Baccalaureate

Comparability of Scores on New Forms of a Cognitive Reasoning Test Internationally
Thanos Patelis, Johns Hopkins U & University of Kansas

Utilizing Web Monitoring and Data Forensics in International Certification Testing
Cicek Svensson, Caveon Test Security

Discussant:
Stephen G Sireci, **University of Massachusetts, Amherst**

065. Clustered eBoard Session 1

9:50 to 11:20 am
Marriott: Floor 7th - Salon I

065-1. Clustered eBoard - Alternate Assessment Participation Electronic Board Session

Participants:

The Relationship between Student Placement and AA-AAAS Participation Rates

Sheryl Lazarus, National Center on Educational; Mari Quanbeck, University of Minnesota - Twin Cities

States struggle to meet ESSA's 1% cap on student participation in alternate assessments based on alternate academic achievement standards (AA-AAAS). This paper examines placement rates in more restricted educational settings and participation in AA-AAAS, hypothesizing that higher rates of segregated learning are correlated with higher AA-AAAS participation rates.

A Longitudinal View of States Meeting the 1.0% AA-AAAS Participation Requirement

Yi-Chen Wu, University of Minnesota/NCEO; Martha Thurlow, National Center on Educational

The Every Student Succeeds Act (2015) placed a 1.0% cap on state-level participation in the alternate assessment based on alternate academic achievement standards (AA-AAAS). This study explored the variability in rates across and within states and examined how AA-AAAS participation rates to explore how participation rates changed over time.

065-2. Clustered eBoard - Automatic Essay Scoring Electronic Board Session

Participants:

Identifying Predictors of Middle School Students' Perceptions of Automated Writing Evaluation

Fan Zhang, University of Delaware; Joshua Wilson, University of Delaware; Tania Cruz, University of Delaware; Corey Palermo, Measurement Incorporated; Halley Eacker, Measurement Incorporated; Matthew Myers; Jessica Coles, Measurement Incorporated; Andrew Potter, University of Delaware

We examined predictors of middle school students' perceptions of an Automated Writing Evaluation system—MI Write. Students' limited-English proficiency status, family income, grade level, classroom climate perceptions, liking writing and recursive process beliefs, and MI Write scores in fall and spring predicted usability, usefulness, and desirability in a differential manner.

Diagnostic Bias in Writing Assessment. Implications for Fair Educational Decisions.

Michael Matta, University of Houston; Sterett H. Mercer, The University of British Columbia; Milena A. Keller-Margulis, University of Houston

This study examines the diagnostic bias of automated writing scores. Findings will show whether automated and hand-rated scoring approaches lead to biased scores against minoritized students. We will also evaluate whether less authentic writing indicators (e.g., multiple-choice questions) negatively impact the measurement of writing performance for Black and Hispanic students.

Impact of Scaling on Automated Essay Scoring

YoungKoung Kim, College Board; Tim Moses, College Board; Luz Bay, College Board

The present study examined the impact of scaling for machine scores to human score scales in automated essay scoring (AES). Two scaling methods – mean/sigma and cubic transformation scaling – were examined in the AES model. The results showed that the AES models with scaling greatly improved prediction accuracy.

065-3. Clustered eBoard - Calibrating Field Test Items in Adaptive Tests Electronic Board Session

Participants:

A New Method to Calibrate Pretest Items in Multistage Adaptive Testing

TsungHan Ho, ETS

Pretest items in multistage tests typically have volumes adequate to produce stable parameter estimates, but the linking item volume fluctuations bring into question the quality of link to the reference scale. The performances of four calibration/linking methods that appeared to be robust to volume variations are evaluated using simulation data.

Impact of Ability Range Restriction on Item Characteristics in Multistage Adaptive Testing

Kyoungwon Lee Bishop, WIDA at University of Wisconsin Madison; Hacer Karamese, WIDA at University of Wisconsin; Xin (Grace) Li, University of Wisconsin-Madison; Yoon Ah Song, Center for Applied Linguistics

This study explores how differences in test takers' ability range affect item characteristics such as item difficulty, point-measure correlation, and fit statistics in multistage adaptive testing (MST) under the Rasch model. Depending on how we administer new field test (FT) items item placement might affect item characteristics in the calibration.

065-4. Clustered eBoard - CAT: Design and Development
Electronic Board Session

Participants:

Effects of Information and Difficulty on Adaptive Test-taking Experience and Performance

Teresa Ober, Educational Testing Service (ETS); Junfan Zhou, Hong Kong Polytechnic University; Jasmine Collard, University of Notre Dame; Ying Cheng, University of Notre Dame

Could differences in design and administration of computerized adaptive tests (CATs) influence test-takers' experience and performance? Differences studied included information about the adaptiveness (informed v. uninformed) and difficulty (b-matching=.50 v. .75). Results suggested no main or interaction effects of conditions on test anxiety, perceived effort or difficulty, engagement, or performance.

Investigating Conditional Adaptivity for Targeting Item Development

Kristin M. Morrison, Curriculum Associates; Kevin Cappaert, Curriculum Associates

This study examines the use of conditional adaptivity indices (Reckase, Ju, & Kim, 2018; Wyse & MacBride, 2021) in an operational interim CAT to provide additional feedback to content designers. This feedback can help to target item development for specific areas of the scale to improve student proficiency estimation.

065-5. Clustered eBoard - CAT Item Selection
Electronic Board Session

Participants:

A Maximin Information Criterion for Item Selection in Computerized Adaptive Testing

Jyun-Hong Chen, National Cheng Kung University; Hsiu-Yi Chao, National Taiwan Ocean University

MaxiMin Information (MMI) criterion is proposed for item selection in CAT. MMI can ensure not only the amount of information obtained but the increase in selected items' discrimination as test progresses. According to simulation, MMI outperformed the other ISRs in terms of comparable RMSE and more balanced item pool usage.

Exploring Parameter Invariance for Adaptively Assessing Reading among Students with Learning Differences

Wanjing Anya Ma, Stanford University; Amy Burkhardt, Stanford University; Jason Yeatman, Stanford University

The study examines the degree of generalizability of an adaptive reading assessment for diverse student groups. We found selecting items by the staircase algorithm is more efficient than the maximum fisher information to recover ability estimates, especially when the parameter invariance holds less consistently among students with language-based learning differences.

065-6. Clustered eBoard - CAT: Scoring
Electronic Board Session

Participants:

Does Change in Performance During a Test Event Justify Weighted Scoring?

Steven Wise, NWEA; G. Gage Kingsbury; Meredith Langi, NWEA

Performance change can occur during achievement test events due to the presence of generalized disengagement factors (e.g., decreasing motivation, fatigue, increased anxiety). This paper investigated an innovative scoring method that reduces the weight given to portions of test events where disengagement was detected. Evidence was shown for improved score validity.

Scoring Method in CAT for Interim Score Estimation, EAP or MLE?

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

This study investigates the MLE and EAP scoring methods in the interim scoring for a fixed-length computerized adaptive tests (CAT). Results will provide information on which scoring method in interim score estimation is preferred given the test length and the ability level of examinees in practice.

Comparing Performances of EAP Scoring Methods in Multistage Testing

Hyung Jin Kim, University of Iowa; Won-Chan Lee, University of Iowa

In pursuit of improving performances of EAP scoring, this study considered three schemes for assigning priors: population-wise, path-wise, and person-wise. Furthermore, the study considered various options for prior distribution, and difference in mean between prior and examinees' abilities, and compared performances of EAP and alternative scoring methods in MST.

065-7. Clustered eBoard - CDM
Electronic Board Session

Participants:

Combining Gibbs Sampling with Hamiltonian Monte Carlo for Bayesian Diagnostic Model Estimation

Alfonso Martinez, University of Iowa; Farhan Niazi, University of Iowa; Jihong Zhang, University of Iowa; Jonathan Templin, University of Iowa

We propose a synthesis of Gibbs sampling and Hamiltonian Monte Carlo (GS-HMC) for estimating Bayesian Diagnostic Models (B-DMs). A simulation study explores the efficacy of the GS-HMC algorithm in parameter recovery and compares the algorithm to current B-DM estimation techniques in terms of computational efficiency, chain autocorrelation, and convergence rates.

Sensitivity of Q-Matrix Verification to Different Types of Q-matrix Misspecification for DCMs
Olga Kunina-Habenicht

This simulation study investigates the robustness of the Q-matrix verification procedure for LCDMs for three types of Q-Matrix misspecification (underspecified, overspecified, balanced) based on a previous study. Under conditions with correct, underestimated, and overestimated Q-matrices the Q-matrix was correctly recovered in almost all cases. For balanced misspecification many errors occurred.

Benefits of Using Cognitive Models Within a Mathematics Large-Scale Assessment
Philipp Sonnleitner, Luxembourg Centre for Educatio; Michael Andreas Michels; Pamela Isabel Inostroza Fernández, University of Luxemburg; Caroline Hornung, University of Luxemburg; Sylvie Gamo, University of Luxemburg

Using cognitive models for item development was often called for due to obvious advantages, but largely ignored by educational large-scale assessments. Based on the Luxembourgish school monitoring program, we demonstrate that investing the effort pays off by higher efficiency through automatic item generation and enhanced feedback through Diagnostic Classification Models.

065-8. Clustered eBoard - Classification Consistency and Accuracy Electronic Board Session

Participants:

A Comparison of Methods to Evaluate the Consistency of Cutscore Decisions
Jordan Nelson Stoeger, Data Recognition Corporation; William Skorupski, Data Recognition Corporation

Decision consistency, or the reliability of pass/fail decisions on an assessment, is a critical measure of the psychometric quality of an assessment. The present study investigates the relative performance of six decision consistency methods applied to simulated data based upon actual administration data from a medical specialty board.

Measuring Consistency and Accuracy of Summative Determinations from a Diagnostic Assessment System
Jordan M. Wheeler, University of Georgia; Laine Bradshaw, Pearson; Madeline Schellman

State accountability assessment systems must establish annual summative determinations to categorize students based on their performance. In this study, we propose a method for estimating measures of classification consistency and accuracy for summative determinations obtained from a diagnostic assessment system. Additionally, the proposed method is applied to empirical data.

Evaluating Classification Accuracy of Screener Assessments with Receiver Operating Characteristic Curve Analysis
Ye Yuan, University of Georgia

This study investigates the classification accuracy of a screener assessment through the Receiver Operating Characteristic (ROC) curve using various cut scores on statewide assessments as criterion measures. The study examines ROC seasonal trends and different cut scores in different groups. The study also explores a variety of predictors of performance.

065-9. Clustered eBoard - Growth 1 Electronic Board Session

Participants:

Growth Modeling for Learning Loss and Recovery: Evidence from a Statewide Assessment
Yen Vo, University of Iowa; Annette Vernon, University of Iowa; Stephen Dunbar, University of Iowa; Catherine Welch, University of Iowa

This study explores how student growth occurred before, during, and after the COVID-19 pandemic by analyzing student growth percentiles across school districts in one midwestern state. Findings showed that growth patterns varied among school districts and considered possible explanations for the observed variation.

Applying Growth Models to Contextualize Longitudinal Performance on Practice Tests
Siyu Wan, ABIM; Francis O'Donnell, National Board of Medical Examiners; Lisa Keller, University of Massachusetts

Little is known about how much progress examinees make across practice tests for high-stakes assessments. We used a dataset from a practice test for a medical licensure exam to explore this topic, applying growth models to total and content area scores. Approaches for contextualizing longitudinal practice test results are discussed.

065-10. Clustered eBoard - Growth 2 Electronic Board Session

Participants:

Confidence Interval of Effect Size in Longitudinal Growth Models
Zonggui Li, Boston College; Ehri Ryu, Boston College

In this study, we aim to provide appropriate confidence interval (CI) computation for the previously developed-squared in longitudinal growth models. Three bootstrap methods and five CI computation methods with complex level-1 residual covariance structure will be examined using simulated data.

065-11. Clustered eBoard - IRT Model Fit
Electronic Board Session

Participants:

The Accuracy of Bayesian Model Fit Indices for MIRT Model Comparison
Ken Fujimoto, Loyola University Chicago; Carl Falk, McGill University

We examined how much the greater fit propensity of certain item response theory models (e.g., the trifactor model) can bias four Bayesian predictive performance indices in their favor when performing model comparisons. The deviance information criterion was the most biased, and a leave-one-out cross-validation approximation was the least biased.

Bifactor Item Response Analysis of the Objective Structured Clinical Examination
Nai-En Tang; Chia-Lin Tsai; Igor Himelfarb

The three competing Grade Response Models (GRMs) (e.g., unidimensional, multidimensional, and bifactor) were analyzed for the chiropractic case management Objective Structured Clinical Examinations license exam. The results show that the bifactor GRM fits the exam well and confirms that there is more than one dimension for the exam.

065-12. Clustered eBoard - Item Parameter Drift
Electronic Board Session

Participants:

Robustness of Four Rasch Banking Procedures When Item Parameters Drift
Chunyan Liu, National Board of Medical Examiners; Daniel Jurich, National Board of Medical Examiners; Peter Baldwin, National Board of Medical Examiners; Wenli Ouyang, National Board of Medical Examiners; Raja G Subhiyah, National Board of Medical Examiners

Many testing programs rely on a large item bank with known pre-equated item parameter estimates. One threat to the integrity of an item bank is item parameter drift and, in this context, we compare the robustness of four item banking procedures and their effects on model parameter estimates over time.

An Experimental Design to Investigate Item Parameter Drift on a Licensure Exam
Peter Baldwin, National Board of Medical Examiners; Irina Grabovsky, 3750 Market Street; Brian Clauser, National Board of Medical Examiners; Kimberly Swygert, National Board of Medical Examiners; Thomas Fogle, National Board of Medical Examiners

A multiyear study investigating unexpected score increases on a high-stakes licensure exam is described. As every exposed item was at risk for item parameter drift, it was decided to test for drift using only unexposed items deployed in a stratified random method within an experimental design. Results are given.

065-13. Clustered eBoard - Missing Responses
Electronic Board Session

Participants:

The Effect of Missing Responses on Bayesian Estimation of Three-Parameter IRT Models
Tzu-Chun Kuo, Kaplan North America; Yanyan Sheng, University of Chicago

This study compared four Bayesian approaches for estimating three-parameter IRT models when data were complete or missing under different types (MCAR, MAR, and MNAR). Preliminary results suggested that the accuracy of parameter recovery was reduced when more missing data were present and/or the missingness was not at random.

Handling Missing Item Responses from Omitted and Not-reached Items in Large-scale Assessment
Dongwei Wang, UMass Amherst; Craig Wells, UMass Amherst

Significant missing responses are observed in large-scale assessments and vary across countries. This study demonstrates the effects of scaling item responses with a model-based approach handling missing responses from omitted and not-reached items simultaneously. Preliminary results showed a meaningful difference in item difficulty when different methods were used for scaling.

065-14. Clustered eBoard - MST Routing
Electronic Board Session

Participants:

Impact of Misrouting in an MST
Jing Ma, The University of Iowa; Xi Wang; Anthony D. Fina, University of Iowa; Catherine Welch, University of Iowa

Ideally, it should not matter what test form a student receives or path a student follows in a multistage test. This study explicitly examines if that holds true in practice by simulating test taker performance on every route. The impact of misrouting on measurement precision and proficiency determinations are summarized.

Investigating the Impact of Misrouting in Multistage Adaptive Testing

Hacer Karamese, WIDA at University of Wisconsin-Madison; Won-Chan Lee, University of Iowa

This simulation study aims to investigate the potential impact of misrouting on MST scores. Simulations are performed to manipulate the panel design and final scoring method to evaluate their effect on the scores. The results and discussion of the findings provide insights into the practical implications of misrouting.

Impact of Routing Methods on Measurement Precision and Error in Multistage Testing

Xi Wang; Jing Ma, The University of Iowa; Anthony D. Fina, University of Iowa; Catherine Welch, University of Iowa

This study examines the impact of two different estimators when the defined population intervals method is used for routing in an MST. Specifically, simulations are used to examine how number correct scores and EAP estimates compare and their impact on classification accuracy, error, and module exposure.

065-15. Clustered eBoard - Non-Cognitive Models Electronic Board Session

Participants:

Maximum Marginal Likelihood Estimation of the MUPP-GGUM Model

Jianbin Fu, Educational Testing Service; Xuan (Adele) Tan, ETS

The Multi-Unidimensional Pairwise Preference with Generalized Graded Unfolding Model (MUPP-GGUM; Stark et al., 2005) is used to calibrate blocks with two ideal-point statements in a forced-choice questionnaire. The current study develops a maximum marginal likelihood estimation with an expectation-maximization algorithm to estimate item parameters and their standard errors of MUPP-GGUM.

Investigating Sense of Belonging in Graduate Students Using Explanatory Item Response Models

Carlos Chavez, University of Minnesota - Twin Cities; Tai Do, University of Minnesota Depart; Michael C. Rodriguez, University of Minnesota

This paper investigates the effect of nested data structure on polytomous item responses using a multilevel explanatory item response model. We raise questions regarding the extent to which item responses vary by persons and by institutions, as well as the impact of major study of area on the category thresholds.

Introducing Network Analysis for Meaningful Country and Group Comparisons in Large-Scale Assessments

Guher Gorgun, University of Alberta; Sevilay Kilmen

Introducing the psychometric network analysis method, we compared the functioning of non-cognitive variables in an international large-scale assessment for students with different immigration status across two different country contexts. The relationship among non-cognitive variables differentiated across country and immigration status. Network analysis can be used for justifying inferences made across countries.

065-16. Clustered eBoard - Perspectives and Methods on Score Reliability Electronic Board Session

Participants:

Estimating Reliability for Tests with One Constructed Response Item in a Section

Yanxuan Qu, ETS; Sandip Sinharay, Educational Testing Service

Educational tests often have sections that are not parallel. In addition, some tests occasionally have only one item in a section. It is unclear how to estimate reliability for these tests. In this study, we propose a two-step approach for estimating reliability in such situations.

Examination of Test Characteristics Effect on Coefficient α and Coefficient ω

Terry Ackerman, University of Iowa; Richard Melvin Luecht, University of North Carolina at Greensboro; Cheryl Ma, Amazon Web Services (AMS)

In this study, five factors (number of items, level of item discrimination, number of dimensions, correlations **between** dimensions, location of latent ability distribution) were simulated to determine their effect on three measures of reliability: α, ω, and true scale reliability. In higher dimensionality conditions ω was significantly lower than α.

065-17. Clustered eBoard - Score Prediction Electronic Board Session

Participants:

Evaluating Methods for Predicting Summative Assessment Performance from Interim Assessment Results

Luciana Cancado, Curriculum Associates; Logan Rome, Curriculum Associates

Educators often want to use interim assessment results to project students' performance on summative assessments. This study compares and cross-validates traditional equipercentile equating to both conventional statistical and machine learning methods for creating predictive relationships between these assessments. We find that the equipercentile method produces results with near optimal accuracy.

Projecting Validity and Reliability for a Shortened Unidimensional Assessment

Joshua Moskowitz, Altus Assessments; Wei Wei Yan, Altus Assessments; Jordan L Ho, Altus Assessments; Colleen Robb, Altus Assessments; Alexander MacIntosh, Altus Assessments; Gill Sitarenios, Altus Assessments

We demonstrate a methodology to project the impact to reliability and validity from shortening a unidimensional assessment. Using historical test data, we apply this methodology to the high-stakes situational judgment test Casper. This methodology can be applied to other unidimensional tests for which historical test data exists.

065-18. Clustered eBoard - Student Testing Behavior
Electronic Board Session

Participants:

Using Automated Item Generation to Provide Individualized Feedback in Formative Tests
Ayfer Sayin; Mark Gierl, University of Alberta
The purpose of our paper is to describe and demonstrate a new method for providing (i) automated, (ii) individualized, and (iii) detailed feedback in computer-based formative testing by extending the item modelling process in the automatic item generation (AIG). The feedback system is demonstrated using two selected-response examples.

Process Data Insights into Effects of Nudges on Cognitively Disengaged Student Behavior
Burcu Arslan, Educational Testing Service Global B.V.; Bridgid Finn, Educational Testing Service
The detection and treatment of disengaged test-taker responses can improve test validity because such responses do not represent test-taker knowledge. We investigated the effect of behavioral nudges, presented whenever student response times were below a predefined item-time threshold, on disengaged-student behavior. Experimental findings and their implications will be presented.

065-19. Clustered eBoard - Using Process Data to Understand Student Behavior
Electronic Board Session

Participants:

Understanding the Student Intercepts in Logistic Knowledge Tracing Models for Measuring Proficiency
Guoguo Zheng, Amplify Education; Reginald Ziedzor; Seyfullah Tingir, Amplify Education; Wanchen Chang, Cambium Assessment
This study investigates how to interpret the student intercepts in logistic knowledge tracing (LKT) models that might be used to measure students' real-time proficiency as they complete practice problems. It also examines the validity of the students' real-time proficiency measured by LKT using empirical learning data.

Using Random Walks to Cluster Answer Change Patterns on Interactive Tasks
Xin Qiao, Southern Methodist University; Tracy Sweet, University of Maryland College Park
We defined answer changes on interactive tasks using a social network analysis method called random walks and investigated answer change patterns in the clickstream data. We conducted cluster analysis on random walk counts and response times. Results revealed different answer change patterns relating to score categories and person characteristics.

Examining Fairness in Joint Process-Response Models Using Person-Fit Statistics
Matthew David Naveiras, Peabody College of Vanderbilt; Xiang Liu, Educational Testing Service; Michael Fauss, ETS
This paper presents the derivation of a joint process-response person-fit statistic to identify poor person fit regarding both item responses and test-taking behaviors. The type-I error rate and power are examined through simulations. A NAEP data set consisting of item responses and calculator usage is analyzed to demonstrate its utility.

066. Supporting Test Security of Remote Testing with Process Data Analytics and AI
Coordinated Paper Session
11:40 to 1:10 pm
Marriott: Floor 5th - Chicago Ballroom B/C

COVID pandemic accelerated the implementation of remotely proctored high-stake assessments. Complementary to traditional psychometric and statistical analyses on the scores, clickstream process data that captures the fine-grained interaction between test takers and items can provide rich diagnostic information to help improve test security. In this coordinated session, we gathered four presentations to show how to apply data analytics and AI techniques to clickstream data to identify imposters in writing, uncover copy writing behaviors, capture speech responses from common reading templates, and detect AI-generated essays. These four examples are part of our comprehensive research agenda on test security, by which we would like to share with the community our efforts to ensure the validity, reliability and fairness of remote testing.

Session Organizer:
Jiangang Hao, **Educational Testing Service**

Chair:
Jiangang Hao, **Educational Testing Service**

Participants:

Benchmark Keystroke Biometrics Accuracy from High-Stakes Writing Tasks
Ikkyu Choi; Jiangang Hao, Educational Testing Service; Paul Deane, ETS; Mo Zhang, Educational Testing Service

Detection of Retyping vs. Drafting Through AI-based Methods based on Keystroke Process Data
Mo Zhang, Educational Testing Service; Paul Deane, ETS; Jiangang Hao, Educational Testing Service

Speech Response Similarity Detection in Remote Testing
Michael Fauss, ETS; Jiangang Hao, Educational Testing Service

Implications of AI-generated Essays
Duanli Yan, ETS; Michael Fauss, ETS; Wenju Cui, ETS; Jiangang Hao, Educational Testing Service

Discussant:
Ye Tong, **National Board of Medical Examiners**

067. Measuring Change in a Changing World: Updating Frameworks without Breaking Trends

Coordinated Paper Session
11:40 to 1:10 pm
Marriott: Floor 5th - Chicago Ballroom D

Educational and professional standards change as science and society advance. When assessment governing bodies respond by adapting content frameworks, they must navigate political and technical challenges, including communicating and justifying the change to the field and evaluating whether score comparisons between the old and new assessment are appropriate. When a primary purpose of the assessment is to monitor progress over time, there appears to be a direct tradeoff between updating content frameworks and reporting meaningful educational and professional progress. In this symposium, Presenters will characterize these tensions and provide political and technical strategies for governing bodies to navigate them. The Presenters have all been commissioned to contribute papers on this topic to inform future updates of content frameworks for the National Assessment of Educational Progress (NAEP). The Presenters will discuss the challenge of measuring progress while updating frameworks in the NAEP context. They will also incorporate examples and strategies motivated by framework updates in other educational and professional contexts. Topics include the role of "bridge studies," whether some subject area domains are easier to update while maintaining trends than others, and the importance of the political context that may motivate the framework update or the preservation of trends.

Session Organizer:
Andrew Ho, **Harvard Graduate School of Education**

Chair:
Sharyn Rosenberg, **NAGB**

Participants:
NAEP Science: Tensions Between Maintaining Trend and Aligning with New Standards
Alicia Alonzo, Michigan State University
NAEP Framework and Trend Considerations
Lorrie Ann Shepard, University of Colorado Boulder
Keeping NAEP Relevant: Considerations for More Frequent Changes to NAEP Assessment Frameworks
Stanley N Rabinowitz, EdMetric LLC
Assessing and Reporting Clinical Judgment Skills in Nursing – A Workforce Example
Ada Woo, Ascend Learning, Christine Mills, Ascend Learning

Discussant:
Andrew Ho, **Harvard Graduate School of Education**

068. Monitoring Performance of U.S. Students in the Pandemic with NAEP Long-Term Trend Assessments

Coordinated Paper Session
11:40 to 1:10 pm
Marriott: Floor 5th - Chicago Ballroom E

Educators and policy makers have been grappling with understanding the impact of pandemic-related interruptions on student learning. The National Assessment of Educational Progress (NAEP), administered by NCES, has been monitoring student performance in mathematics and reading through age-based long-term trend (LTT) assessments since the early 1970s. This coordinated paper session brings together NCES and various institutions to discuss the 2022 age 9 NAEP LTT data collection and analysis efforts to contribute to the conversation on the impact of pandemic-related interruptions on academic progress. The collection of 4 papers will address the importance of the 2022 age 9 NAEP LTT assessment data collection. We will discuss sampling, data collection and analysis in providing critical education statistics on the U.S. age 9 student population and student groups in 2022, as well as in monitoring changes in student performance between 2022 and previous years. In addition, the section contains research that delves deeper in exploring relationships between changes in performance and other covariates, including demographics, background variables and instruction mode.

Session Organizer:
Yue Jia, **Educational Testing Service**

Chairs:
Yue Jia, **Educational Testing Service**
Amy Dresher, **ETS**

Participants:

The Importance of the 2022 NAEP LTT Assessment
Amy Dresher, ETS
2022 Age 9 NAEP LTT: Sample Design, Population Coverage, and Participation Rates
Leslie Wallace, Westat; Keith Rust, Westat
Analysis of 2022 Age 9 NAEP LTT and Trend Comparison
Adrienne Sgammato, ETS; Nuo Xi, Educational Testing Service
Exploring Impact of Pandemic-Related Disruptions on Student Performance: Age 9 NAEP LTT
Katherine Furgol Castellano, Educational Testing Service; Daniel McCaffrey, Educational Testing Service; Nuo Xi, Educational Testing Service; Yue Jia, Educational Testing Service; Laura Hamilton, American Institutes for Research

Discussant:
Derek Christian Briggs, **University of Colorado Boulder**

069. Measurement Models for the Purpose of Evaluating Interventions and Programs

Coordinated Paper Session
11:40 to 1:10 pm
Marriott: Floor 5th - Chicago Ballroom F

Though much effort is often put into designing studies, the measurement model and scoring approach employed are often an afterthought, especially when short survey scales are used (Flake & Fried, 2020). One possible reason that measurement gets downplayed is that there is generally little understanding of how calibration/scoring approaches could impact common estimands of interest, including treatment effect estimates, beyond random noise due to measurement error. Another possible reason is that the process of scoring is complicated, involving selecting a suitable measurement model, calibrating its parameters, then deciding how to generate a score, all steps that occur before the score is even used to examine the desired psychological or educational phenomenon. In this series of studies, we propose IRT models specifically developed to match four study designs: (1) multisite cluster RCTs, (2) difference-in-difference, (3) regression discontinuity, (4) teacher observations using multi-rater protocols. We show that using a measurement model matching the study design often substantially reduces bias in estimated treatment effects, Type I and II error rates, and misclassification error relative to scoring in a way that does not match the design.

Session Organizer:
James Soland, **University of Virginia**

Participants:

Measurement Models in the Context of Multi-site Cluster RCTs
Megan Kuhfeld, NWEA; James Soland, University of Virginia

In this study, we conduct a series of simulation studies that consider a wide range of options for producing scores in the context of multisite RCTs. We find that the true treatment effect is attenuated when scores from IRT models that do not account for treatment assignment are used.

Measurement Models in the Context of Difference-in-Difference Designs
James Soland, University of Virginia

We present several possible IRT models that match various difference-in-difference (DiD) specifications, describe the potential benefits of using an item-level approach to DID estimation, conduct brief simulation studies examining the performance of such IRT models compared to more traditional ways of producing DiD estimates, then discuss implications.

Improving the Utilities of Regression Discontinuity Analysis by Reinstating Measurement Models
Yang Liu, University of Maryland, College Park; Monica Morrell, University of Maryland; Ji Seung Yang, University of Maryland; Youngjin Han, University of Maryland College Park

This presentation will illustrate how the regression discontinuity (RD) analysis can be benefitted by integrating measurement models into the analysis. Definition of causal estimands, model specification, and consideration of estimation issues are discussed along with the proof-of-concept simulation study and an empirical data motivated but simulated data example.

Measurement Models in the Classroom Observational Protocols
Kelly Edwards, University of Virginia; James Soland, University of Virginia

Although many states have incorporated classroom observations into their teacher evaluation systems, scores are often affected by construct-irrelevant sources of variance like the rater, items, and the lesson being taught. In this study, we present an IRT-based approach to scoring observational protocols that accounts for these sources of construct-irrelevant variance.

Discussant:
Li Cai, **UCLA**

070. Meeting the Challenge: The Law School Admission Test in Changing Times

Coordinated Paper Session
11:40 to 1:10 pm
Marriott: Floor 5th - Chicago Ballroom G/H

The Law School Admission Test (LSAT) has undergone substantial transitions in recent years. Beginning with the July 2019 administration, the LSAT transitioned to the Digital LSAT, a tablet-based test which replaced the paper-and-pencil (P&P) administration mode, yet retained the event-based, test center model of P&P testing. Just eight months later, growing concern over the COVID-19 pandemic would render test centers inaccessible. In response to this unprecedented disruption, the Law School Admission Council (LSAC) launched LSAT-Flex in May 2020, a shortened (three section) version of the LSAT that could be delivered online with remote proctoring. However, operational constraints at that time necessitated the suspension of item pretesting. Eventually, the moratorium on pretesting was lifted, and in August 2021, LSAC transitioned to a four-section LSAT which included three operational sections (as in LSAT-Flex) but reintroduced the unscored pretest section. This session will highlight the key transitions in the LSAT testing program over the past few years, outlining the research studies, designs, and analyses for supporting score interpretations at each stage.

Session Organizer:
Anna Topczewski, **Law School Admission Council**

Chair:
Janeen McCullough, **Law School Admission Council**

Participants:

The Comparability Between Paper-Based and Digital-Based LSAT Scores
Aolin Xie, Law School Admission Council
This study examined score comparability between the paper-and-pencil and digital modes of the LSAT using post-administration testing data. Both item-level and test-level statistics were evaluated for mode effects. The comparability evidence collected ensured that LSAT scores from the two administration modes could be interpreted similarly.

The LSAT: Practical Considerations of Transitioning to Online Remote Proctored Testing
Josiah Evans, Law School Admission Council
In response to COVID-19, the LSAT transitioned from electronic tablets to online remote-proctored testing. The transition posed significant challenges related to operations, validity, fairness, access, and selection of a remote-proctoring vendor. In this presentation you will learn about LSAC's solutions to these daunting challenges.

Transitioning from a Five-Section LSAT to a Three-Section LSAT-Flex
Yu Fang, Law School Admission Council
During the COVID-19 crisis, a shortened version of the LSAT, "LSAT-Flex," was launched. This study examined the impact of shortened tests on the test statistics and equating results based on the historical LSAT data and evaluated score trends with empirical LSAT-Flex data.

Examining LSAT Score Comparability After Reintroducing Pretest Sections
Anna Topczewski, Law School Admission Council; Janeen McCullough, Law School Admission Council
Due to COVID-19, LSAC quickly pivoted to an online remote-proctored three-section test ("LSAT-Flex") in May 2020. With LSAT-Flex, pretesting had been suspended for over a year, until LSAC reintroduced a fourth unscored section in August 2021. This paper focuses on the comparability study conducted.

Discussant:
Alexander Weissman, **Law School Admission Council**

071. Culturally Responsive and Related Approaches to Assessment: What are They?

Organized Discussion
11:40 to 1:10 pm
Marriott: Floor 5th - Denver/Houston

The growing recognition of the need to address longstanding issues of social justice is causing some state education departments to revise content standards and some educational assessment organizations and many members of our field to rethink traditional definitions of equity and fairness. Evolving definitions include in them the need to account for differences in sociocultural backgrounds, funds of knowledge, interests, values, and practices that individuals from diverse cultures bring to learning and assessment. In keeping with this need, a variety of approaches to assessment have been proposed. Among those approaches are culturally responsive assessment, socioculturally responsive assessment, antiracist assessment, culturally sustaining assessment, justice-oriented assessment, and universal design for assessment. The goal of this session is to facilitate an organized discussion that helps to clarify these concepts. Among the questions the panel will address are how these approaches are similar, how they differ, what students populations are put in focus, how the approaches are being (or might be implemented) in practice, and how one might best communicate about them.

Session Organizer:
Randy Bennett, **ETS**

Presenters:
Molly Faulkner-Bond, **WestEd**
Guillermo Solano-Flores, **Stanford University**
Aneesha Badrinarayan, **Learning Policy Institute**
Leanne Ketterlin Geller, **Southern Methodist University**

072. Assessing the Impact of Feedback in Computer-Based Assessments

Paper Session
11:40 to 1:10 pm
Marriott: Floor 5th - Los Angeles/Miami

Chair:
Robert Schwartz, **ACT**

Participants:

Feedback in Computer-Based Assessments: Effects on Test-Taker Affect and Performance
Ute Mertens, Leibniz Institute for Science and Mathematics Education; Marlit Annalena Lindner, IPN Kiel

We experimentally varied automated feedback (Elaborated Feedback vs. Answer-Until-Correct vs. no Feedback) to investigate effects on students' affect during a computer-based science test. Item-level performance-related effects on test-takers' affective states emerged (positive effects following correct answers; negative effects following incorrect answers). However, elaborated feedback mitigated negative emotions after incorrect answers.

Investigating How Emotional Affect Moderates the Relationship Between Feedback Type and Uptake
Samuel Dale Ihlenfeldt, University of Minnesota; Joseph A. Rios, University of Minnesota

This study explores the relationship between feedback presentation, emotional affect, and feedback uptake. Participants were randomly assigned to one of three feedback conditions after taking a GRE practice test. Results indicate that both affect (positive and negative) and feedback condition are significantly associated with reported likelihood to change studying behaviors.

Feedback in Computer-Based Assessment: Does Adding Representational Pictures and Emotional Design Matter?
Livia Kuklick, IPN Kiel, Germany; Marlit Lindner, IPN Kiel, Germany

While computer-based assessments facilitate the implementation of automated text-based and multimedia feedback, it is unclear whether adding visual design features to feedback messages enhance their benefit. This experiment examines the cognitive and affective effects of adding a representational picture or emotional design to immediate feedback messages in a low-stakes assessment.

Test Takers' Desire for Computer-Based Feedback on Low-Stakes Assessments: Insights from Self-Reports
Marlit Lindner, IPN Kiel, Germany; Ute Mertens, Leibniz Institute for Science and Mathematics Education; Livia Kuklick, IPN Kiel, Germany; Christian Schöber, IfBQ Hamburg; Steven Wise, NWEA

This study shows that students seek performance feedback on low-stakes assessments (LSA), even when their results have no personal consequences. Students' desire for feedback was broadly comparable across achievement levels, gender, and school tracks in two independent data sets. Potential and challenges of implementing computer-assisted feedback in LSA are discussed.

Assessment in the Classroom: A Question of Grainsize
Mark Wilson, Berkeley School of Education, UC Berkeley

Educational assessments can be related to two levels of teacher use in the classroom: meso (externally-developed items) and micro (on-the-fly in the classroom). Elementary school measurement assessments and related apps were developed and tried out, and empirical results are described. The benefits of this perspective are discussed.

Discussant:
Bozhidar M. Bashkov, **IXL Learning**

073. GSIC eBoard Session 1

Graduate Electronic Board Session
11:40 to 1:10 pm
Marriott: Floor 7th - Salon I

Participants:

1. **Faculty's Intercultural Teaching Competence in Teaching International Students: A Survey Development Study**
Xiaowen Chen, George Mason University
This paper illustrated the qualitative phases of the development of the survey, which focuses on understanding faculty's intercultural competence from international students' perspectives. The results of the survey could provide a comprehensive understanding of international students' experience. It also provides information for the design of professional development workshops for faculty.

2. **Advancing Rating Quality in Social Science and Educational Research**
Yvette Yvette Jackson; Maria Elena Oliveri, Buros Center for Testing-UNL
A latent trait model is explored to examine rating quality in rater-mediated activities designed to illustrate the alignment of test items to educational standards. Findings demonstrated subject matter expert (SME) scores formed a strong to moderate scale with no significant violations of monotonicity and invariant rater ordering.

3. **Impact of MST Design Decisions on Precision and Error: Lessons for Practice**
Hong Chen, University of Iowa; Ahmed Bediwy, The University of Iowa; Mubarak Olumide Mojoyinola, The University of Iowa; Anthony D. Fina, University of Iowa; Catherine Welch, University of Iowa
Critical design decisions can impact how an MST functions. This study examines how different decisions impact score precision and error in an MST. Specifically, difference in average difficulty (small vs large) between modules, number of modules per stage and number of stages are examined. Practical implications for practice are discussed.

4. **Validation of the German Version of the Receptivity to Instructional Feedback Scale**
Jan Luca Bahr; Lars Höft, Leibniz Institute for Science and Mathematics Education; Jennifer Meyer, Leibniz Institute for Science and Mathematics Education; Thorben Jansen, Leibniz Institute for Science and Mathematics Education
The current study validated a German version of the Receptivity to Instructional Feedback scale. Specifically, the superiority of exploratory structural equitation modeling over traditional confirmatory factor analysis in assessing the construct has been shown. Contrary to expectations, we did not find a global factor to the receptivity construct.

5. **Generalize Bifactor Model Within Partially Confirmatory Factor Analysis**
Yifan Zhang
The partially confirmatory factor analysis (PCFA) has accommodated various bifactor models in standard-type with continuous data. This study will extend this framework to accommodate mixed-type responses and more conditions. Experiments will be conducted to test the recovery performance and advantages of inheritance in standard-type and extended-type bifactor.

6. **A Partially Confirmatory Cognitive Diagnosis Model with Polytomous Data**
Yi Jin
Most cognitive diagnosis models are only conducted for dichotomous response data. To address the issue, this research proposes extending the partially confirmatory cognitive diagnosis model to accommodate polytomous data with the Bayesian Lasso approach. Also, MH-within-Gibbs sampling algorithm would be adopted with an analysis of Markov Chain Monte Carlo simulation.

7. **Impact of Academic and Socioeconomic Factors on University Admission in Chile**
Geraldo Bladimir Padilla; Mubarak Olumide Mojoyinola, The University of Iowa
This paper presents an in-depth quantitative analysis of the impact of academic and socioeconomic factors on the probability of being admitted into university in Chile. Estimating the probability of admission based on these factors will provide some insights into quantifying the extent of inequality in the admission process.

8. **Effects of Sample Size and Collapse Direction on Parameter Recovery**
Yale Quan, University of Washington; Chun Wang, University of Washington
In ordinal data analysis, category collapse occurs when adjacent response options are combined. We extended the current research on using GRM and GPCM with collapsed data by exploring how sample size and direction of collapse influence parameter estimation. Simulation studies will be presented along with an empirical application.

9. **HG-MI Scoring: A Two-stage Ability Estimation Procedure for Handling Rapid Guessing**
Jiayi Deng, University of Minnesota, Twin Cities; Joseph A. Rios, University of Minnesota
This simulation study explored ability estimation accuracy for a multiple imputation extension of the Holman and Glas (2005) model when manipulating: (a) sample size; (b) the covariance between ability and rapid guessing (RG); (c) RG percentage; and (d) RG classification accuracy (both type and percentage).

10. **Addressing Test-Taking Disengagement Across Multiple Domains Simultaneously Using an IRTree Model**
Katarina Schaefer, James Madison University; Brian C Leventhal, James Madison University
In addition to examinee ability, test scores reflect effort which is especially problematic in low-stakes testing. IRTree models have characteristics that traditional methods used to address low-effort do not have (easier-to-meet data assumptions and higher precision). We present the promising results for the use of this model across multiple domains.

11. **Applying Adaptivity Indices for a Large-Scale Interim Assessment Program**
Joselyn Perez; Kristin M. Morrison, Curriculum Associates
The current paper demonstrates the use of four adaptivity indices described by Reckase, Jui, and Kim (2019) and Cui (2020) for an operational interim computer adaptive test (CAT) assessment program. The degree of adaptivity was studied across grades, test season, and repeated testing conditions.

12. **Cognitive Diagnosis Models for Disengaged Behaviors**
Benjamin Kweku Lugu; Wenchao Ma, University of Alabama; Wenjing Guo, University of Alabama; Randall Schumacker
The study will use real and simulated data to examine the performance of cognitive diagnostic model (CDM) integrated with disengaged behavior arising from low-stake tests. The proposed model features missing response, guessing and attempt of problem-solving. The model will be fitted with the G-DINA model.

13. **Comparing Profile Patterns of Mathematics and English in K-12 Testing**
Haeju Lee, University of North Carolina Greenboro; Hongwook Suh, Cambium Assessment, Inc.; Kyung Yong Kim, University of North Carolina Greenboro
The purpose of this study is to investigate the patterns of individual differences in K-12 assessment and to illustrate what variables interact with the math and English performances of groups. Latent profile analysis (LPA) was used to examine the profile patterns and the effect of covariates on performances.

14. **Annotation and Rater Training Guideline for Annotating Quality of Students Argumentative Essays**
Nils-Jonathan Schaller, IPN – Leibniz Institute for Science and Mathematics Education
To approach the lack of datasets focused on student's high school education, we explain our annotation process for a dataset consisting of 4500 German high school students' written argumentation texts and propose annotation guidelines for measuring argumentation quality to provide adaptive feedback.

15. **Classification Consistency and Accuracy Indices for Simple Structure MIRT Model**
Huan Liu, The University of Iowa; Won-Chan Lee, University of Iowa
This study aims to investigate and extend the classification consistency and accuracy indices using the Lee, Rudner, and Guo approaches under the SS-MIRT framework. The preliminary results show that the Lee and Rudner approaches produced comparable classification indices, which were slightly smaller than those produced by the Guo approach.

16. **A Comparison of Standard Error Estimation Methods for Dichotomous and Polytomous Items**
Meghan Leeming, University of North Carolina at Greensboro
This study is a comparison of both dichotomous and polytomous items in unidimensional and multidimensional item response theory modeling. Six different standard error estimation methods will be used to compare the unidimensional and multidimensional models. Bias and root mean squared error will be used to evaluate the comparison.

17. **A Comparison of Attribute Hierarchies in TIMSS Mathematics between Korea and the U.S.**
Sehee Lee; Hyunsook Lee; Minkyoung Shin, Seoul Women's University; Yoonsun Lee, Seoul Women's University
This study aims to compare Korean and American eighth graders' attribute hierarchies in the TIMSS Mathematics assessment. This study uses an explanatory approach of learning attribute hierarchies from the data and it will provide educators with resources to understand cognitive domains and effective instructional implications.

18. **Effect of Ability Distributions on IRT Observed Score Equating under Common-Item Nonequivalent Groups Design**
Min Liang; Won-Chan Lee, University of Iowa
The main purpose of this study was to investigate the influence of ability distributions on the accuracy of IRT observed score equating results under the common-item nonequivalent groups design. A simulation study will be conducted to investigate five study factors: (1) population distribution, (2) IRT model, (3) calibration method, (4) common-item effect size, and (5) sample size.

074. Culturally Responsive Assessment: Method and Impact

Paper Session
1:30 to 2:30 pm
Marriott: Floor 5th - Chicago Ballroom B/C

Chair:
Yong Luo, **NWEA**

Participants:

Experimental Examination of the Impact of Culturally Responsive Assessment Practices
Laine Bradshaw, Pearson; Madeline Schellman
Much of the work in the area of cultural responsiveness in educational assessment has been theoretical or qualitative. This study adds empirical, quantitative results that can help inform the movement toward culturally sustaining assessment practices to contribute to practices that can help educators make better decisions to support student learning.

Culturally Responsive Performance Assessment: Extending Cognitive Labs to Collect Evidence about Meaningfulness
Carla M. Evans, National Center for the Improvement of Educational Assessment
Research around culturally responsive assessment has focused on the importance of performance-based assessment. Yet the ways in which students' cultural/social identities interact with their academic identities is not well known. This study extends previous research by analyzing meaningfulness through extending cognitive labs for about fifty grades 3-12 students in Hawaii.

Understanding Student Responses to Items from a Sociocultural Perspective: Implications for Measurement
Jose R. Palma, The University of Texas at Austin; Doris Luft Baker, The University of Texas at Austin; Tim Andress, The University of Texas at Austin
Consistent with leveraging measurement for better decision, we used a sociocultural framework to understand how Spanish-English Bilingual students respond to items in a science vocabulary test. Evidence of sociocultural contexts that influence how students respond were identified, and interactions with item difficulties. Implications are discussed in the paper.

Discussant:
Brian Gong, **Center for Assessment**

075. Recent Evidence from the Pandemic and Test Optional Admissions Policies?

Coordinated Paper Session
1:30 to 2:30 pm
Marriott: Floor 5th - Chicago Ballroom D

Session Organizer:
Wayne J. Camara, **LSAC**

Chair:
Wayne J. Camara, **LSAC**

Participants:

Test-optional and the Role of Consistent/Discrepant Scores and HSGPA
Ty Cruce, ACT; Edgar Sanchez, ACT

This study examines the potential role that the discrepancy between students' test scores and HSGPA plays in students' decisions to submit their test scores to a test-optional college. We also examine the role of this discrepancy in explaining observed differences in first-year college success between score senders and non-senders.

New Post-Pandemic Evidence on Admissions Trends, Student Score Sending and College Outcomes
Jessica Howell, The College Board

The Admissions Research Consortium (ARC) provides participating colleges with research to understand pandemic-related changes to applications, admissions, and enrollment in the fall 2021 admissions cycle as well as changes to student test score disclosure/withholding behavior under widespread test-optional policies. Evidence on first-year performance and retention outcomes for 2021-22 freshman is also shared.

Longitudinal Trends in US Undergraduate Admissions Policies: Comparisons by Institutional Characteristics
Sugene Cho-Baker, ETS; Harrison Kell, ETS

Using national-level longitudinal data of higher education admissions, the current study explores changes in the consideration of application materials, including standardized tests. We found a general decline in standardized test requirements, coinciding with an increase in secondary school GPA requirements. Diversity has grown across institutions regardless of test requirement policies.

Discussant:
Michael C. Rodriguez, **University of Minnesota**

076. Content-Referenced Growth: Creating Instructionally Actionable Growth Interpretations in Reading and Mathematics Assessments

Coordinated Paper Session
1:30 to 2:30 pm
Marriott: Floor 5th - Chicago Ballroom E

In this session, we continue our exploration of a novel approach to ascribing content-referenced meaning to students' scores using Curriculum Associates' i-Ready Diagnostic. The goal of "content-referenced growth" is to support interpretations of students' scores relative to both the status of their understanding at one point in time, and their growth in understanding across points in time, relative to the content contained in the assessment. We will expand upon the foundations for this approach presented at NCME in 2022 (Wellburg, 2022; Briggs, 2022, Student, 2022) which focused on fractions as an exemplar learning progression for interpreting growth on a vertical scale of mathematics. This coordinated session will include three papers which build upon this work. In the first paper, we will present results from qualitative teacher interviews conducted with an interactive prototype of the fractions learning progression. The second paper will describe additional learning progressions developed in mathematics which address functions, 2-D geometry, and geometric measurement and how these progressions map to the vertical scale. The third paper will describe the development of learning progressions in reading—one in foundational skills and another in reading comprehension—and share unique challenges to working in this domain.

Session Organizer:
Laurie Davis, **Curriculum Associates**

Chair:
Laurie Davis, **Curriculum Associates**

Participants:

Teacher Reactions to an Interactive Prototype of Content Referenced Growth
Derek Christian Briggs, University of Colorado Boulder; Sanford Student, University of Colorado Boulder

Overview of Learning Progressions for Four Big Ideas in Mathematics
Kyla Mcclure, University of Colorado Boulder; Sarah Wellberg, University of Colorado, Boulder

An Overview of the Reading Foundational Skills Learning Progression
Olivia Cox, University of Colorado Boulder

Discussant:
Leslie Keng, **Center for Assessment**

077. Differential Item Functioning Detection Methods

Paper Session
1:30 to 2:30 pm
Marriott: Floor 5th - Chicago Ballroom F

Chair:
Hyeon-Ah Kang, **University of Texas at Austin**

Participants:

Linearizing the Item Characteristic Curve for Detecting Differential Item Functioning
Montserrat B Valdivia Medinaceli, Indiana University Bloomington; Leslie Rutkowski, Indiana University; Dubravka Svetina Valdivia, Indiana University; Sean Joo, University of Kansas

We propose an alternative approach to compute the difference between observed and expected ICCs to detect differential item functioning. This new method addresses issues with the existing method (root mean square deviation; RMSD). The proposed method was evaluated via a simulation study, and the results were compared to the RMSD.

Detecting Differential Item Functioning Among Multiple Groups Using IRT Residual DIF Framework
Hwanggyu Lim, Graduate Management Admission Council; Danqi Zhu; Edison M. Choe, Renaissance; Kyung (Chris) Han, Graduate Management Admission Council

This study introduces an extended version of the IRT residual DIF detection framework (RDIF; Lim et al., 2022a) for multiple groups. A preliminary simulation study demonstrates its potential as a powerful and convenient method to assess DIF among multiple groups simultaneously, preserving the inherent advantages of the RDIF framework.

Detecting DIF in Response Time Using Bootstrap Percentile Method
Qizhou Duan, University of Notre Dame; Ying Cheng, University of Notre Dame

A differential item functioning method for response time utilizing bootstrap percentile interval is proposed. A simulation study will be conducted to evaluate the performance of the method. Additional effect size measures associated with the method will also be investigated.

Discussant:
Scott Monroe, **UMass Amherst**

078. Predicting Item Difficulty of Language Tests

Paper Session
1:30 to 2:30 pm
Marriott: Floor 5th - Chicago Ballroom G/H

Chair:
Polina Harik, **National Board of Medical Examiners**

Participants:

Utilize Deep Language Model to Predict Item Difficulty of Language Proficiency Tests
Jiyun Zu, Educational Testing Service; Ikkyu Choi; Elizabeth Park, Educational Testing Service; Yuan Wang, ETS

To fulfill the greater demands for new items with “known” difficulty, we predict item difficulty using automatically generated features that involve little human judgement. We hypothesize that language models would contribute strongly. We built and studied performance of prediction models for a listening item type from a language proficiency test.

Item Difficulty Modeling for Language Test
Xueming Li, NWEA; Janice Lee Johnson, NWEA

It is challenging for item developers to determine item difficulty. This lack of predictability leaves them dependent on expensive, time-consuming field testing. This study presents and evaluates a method for modeling item difficulty with contextual word embedding and supervised learning for various item types. The results demonstrate strong model performance.

Discussant:
Minju Hong, **University of Arkansas**

079. Re-thinking Construct Definitions and Measurement Methods to Include Black and Hispanic Cultures

Coordinated Paper Session

1:30 to 2:30 pm

Marriott: Floor 5th - Denver/Houston

Recent events in the United States, including the structural inequities highlighted by the coronavirus pandemic and the Black Lives Matter movements, have helped to refocus both education and education measurement to the reality that our system and its assessments do not fully serve all communities. This session presents three examples of such research, all with the aim of considering the role of race/ethnicity in measurement, by examining the construct through the lens of construct definition, test adaptation, and scoring. The first presentation re-examines the construct of kindergarten readiness for non-English speakers. The second presentation examines the adaptation of a middle-school test to reflect contexts specific to Black and Hispanic communities. The third presentation examines the use of African American Vernacular English (AAVE) in student writing, what the construct represents, and examines how that use relates to scores assigned by both human and automated raters. Each Presenters will provide an overview for her research, and then the discussant will facilitate a conversation on the potential implications of the research. The focus of the discussion will be on what further research is needed and how research can inform measurement.

Session Organizer:
Susan Lottridge, **Cambium Assessment, Inc**

Participants:

Rethinking the Construct: What does a non-English Measure of Kindergarten Readiness Entail?
Marianne Perie, WestEd

Investigating Mirror and Window Item Items Reflecting Diverse Populations
Molly Faulkner-Bond, WestEd; Marianne Perie, WestEd; Priya Kannan, WestEd

The Presence of AAVE Linguistic Markers in Student Writing
Jaylin N Nesbitt, James Madison University; Mackenzie Young, Cambium Assessment; Susan Lottridge, Cambium Assessment, Inc

Discussant:
Catherine Close, **Renaissance Learning**

080. Research Blitz: Various Uses of Process Data

Research Blitz Session

1:30 to 2:30 pm

Marriott: Floor 5th - Los Angeles/Miami

Chair:
Brian C Leventhal, **James Madison University**

Participants:

Use NAEP Process Data to Profile Cognitive Strategies for Solving Spatial Problems
Xin Wei, SRI International; Susu Zhang, University of Illinois at Urbana-Champaign; Jihong Zhang, University of Iowa

This study analyzed 2017 NAEP 8th grade process and performance data to profile the cognitive strategies used by students during their interaction with a mental rotation task. Four distinct profiles were identified: internal visualizers (55%), cognitive offloaders (15%), external visualizers (5%), and the non-triers (25%).

Using Process Data to Identify Latent Profile Patterns of Executive Functioning in Preschool Children
*Nixi Wang, University of Washington; Evelyn Law, National University of Singapore; Jane Sum, A*STAR*

Process data opened up possibilities to enhanced measurement. We conducted a test on children's executive functions using iPads. Response time variables are used for a latent profile analysis to delineate profiles of working memory, inhibition, and cognitive flexibility, and explore how individual and contextual variables are related to profile membership.

Screen Visits and Performance in a Computerized TIMSS Math Test
Serap Büyükkıdık, Ohio State University; Paul De Boeck, OSU

Explanatory item response model analyses based on log-files show that frequency of screen visits during a computerized TIMSS math test was negatively related to performance level. We further study whether the relation generalizes across item types and respondents, and whether the relation holds after controlling for item and individual differences.

Exploring Students' Navigational Pathways in NAEP Process Data
Jacob Maibach, University of Arizona; Juanita Hicks, American Institutes for Research

Much research assumes students navigate linearly through standardized tests, but a recent study with NAEP data found only 10% of students move in a fully linear way. Using process data from the 2017 NAEP Grade 4 mathematics assessment, we explore common nonlinear navigational patterns through k-means clustering and LPA.

Classifying Normal and Deviant Test-taking Status through Unsupervised Classifier
Suhwa Han, University of Texas at Austin; Hyeon-Ah Kang, University of Texas at Austin
The study presents a novel application of machine learning unsupervised classifiers to identify an examinee's deviant test-taking behavior within a test. Utilizing measurement derivatives of well-validated cognitive and behavioral indicators, the algorithms allow for automatic identification of an item set on which the examinee exhibits deviant test-taking status.

The Relationship Between Disengagement and the Time of Day That Testing Occurs
Steven Wise, NWEA; Megan Kuhfeld, NWEA; Marlit Annalena Lindner, IPN Kiel
This study investigated the degree to which the time of day that achievement testing occurred was related to the prevalence of test-taking disengagement. Results showed that, as a school day progressed, both rapid guessing and performance change showed meaningful increases. Implications for testing policy and validity are discussed.

081. GSIC eBoard Session 2
Graduate Electronic Board Session

1:30 to 2:30 pm
Marriott: Floor 7th - Salon I

Participants:

1. **A Comparison of Reliability Coefficients for Single-Administration Survey**
Oxana Rosca, University at Albany - SUNY; Kimberly Colvin, University at Albany, SUNY; Heidi L. Andrade, University at Albany; Jason Bryer, City University of New York
This study compares estimates of internal consistency, i.e., α, β, and ω, for an inventory that includes scales with uni- and multi-dimensional factor structures. Results will be discussed in terms of recent literature on reliability estimates that are both appropriate for multidimensional scales and robust to violations of τ-equivalency.

2. **A Feature Analysis Approach to Measuring Cognitive Complexity in Mathematics Problem Solving**
Deborah La Torre, UCLA
Policymakers recognize the need to improve item development and the usability of test scores in state tests. This study explores the use of a feature analysis approach to defining cognitive complexity as well as the use of explanatory item response models and hierarchical diagnostic models to advance these goals.

3. **A Multilevel Framework for DIF Detection in Computerized Adaptive Testing**
Dandan (Danielle) Chen, University of Illinois at Urbana-Champaign; Justin L. Kern, University of Illinois at Urbana-Champaign; David Shin, Pearson; Jinming Zhang, University of Illinois at Urbana-Champaign
A multilevel framework is presented to account for random effects in adaptive item selection in computerized adaptive testing (CAT). Based on this framework, we propose a two-level logistic regression model and examine it in comparison with the traditional logistic regression procedure in accuracy of DIF detection with item-adaptive CAT data.

4. **Analysis of Students' Attitudes Toward Learning on Mathematics Achievement Using PISA 2015**
Kahee Han, University of Kansas
Data from PISA 2015 was used to see the relationship between schoolwork-related anxiety and achievement motivation, and students' mathematics performance. Also, in order to examine the cultural influences on academic anxiety and motivation, the response patterns of students' anxiety and motivation levels were compared by country and performance level.

5. **A Novel Examination of the Validity of None of the Above**
Kathryn Nicole Thompson, James Madison University; Brian C Leventhal, James Madison University
The use of none of the above (NOTA) is considered one of the more controversial item-writing guidelines. This controversy has led to the examination of NOTA as it influences psychometric item properties (i.e., difficulty, discrimination, and reliability). Instead, we focus on NOTA as it influences examinee response processes.

6. **Comparing the Performance of Three Scoring Methods in Longitudinal Certification Assessments**
Hongyu Yang
Many medical boards are replacing traditional one-time certification exams with longitudinal assessments, which span for years and allow examinees to grow during the assessment. This ability growth brings difficulty in scoring and making valid certification decisions. This simulation study compares the performance of three scoring strategies under different growth conditions.

7. **Determining the Impact of Differential Item Functioning at the Test Level**
Heather Leigh Kayton, University of Oxford; Yasmine El Masri, University of Oxford; Victoria A Murphy, University of Oxford
This paper investigates differential functioning across three language versions of PIRLS 2016 in South Africa using an IRT-based likelihood-ratio test approach. Preliminary findings show that despite more than 25% of items functioning differently across language versions, the overall effect at the test level was negligible.

8. **Examining Difference-in-Difference Estimation when the Common Trend Assumption is Violated**
Hongyu Yang
Difference-in-difference estimators assume that, without treatment, the difference between control and treatment groups is constant over time. While in some studies, this assumption would not strictly hold. We will explore how the results be biased by continuing to estimate despite the assumptions that might be violated.

9. **Predictive Modeling on Test Process Data using Supervised ML Models**
   Mingqin Zhang, University of Iowa; Cheryl Ma, Amazon Web Services (AMS); Catherine Welch, University of Iowa
   Process data has drawn attention from researchers and practitioners as it not only exhibits examinees' testing activities but also uncovers behavioral patterns and cognitive strategies of examinees. This study focuses on students' test strategies of speed and explores predictive performance of process data by supervised machine learning (ML) models.
10. **Qualities of Teaching Practice Observation Data: Application of Item Response Theory**
   Francis Ankomah, University of Cape Coast
   This study will assess the qualities of teaching practice observation data from a university in Ghana with the application of item response theory. Through the census method, all data obtained from 1170 student teachers for the 2020/2021 academic year will be used.
11. **The Impact of Q-matrix on Classification Accuracy and Reliability in Longitudinal Diagnostic Classification Models**
   Olasunkanmi Kehinde, Washington State University; Shenghai Dai, Washington State University; Brian French, Washington State University
   The current study investigates the impacts of different Q-matrices specifications on classification accuracy and reliability for the longitudinal cognitive diagnostic model. A simulation study would be performed to investigate factors that might influence the classification accuracy and reliability, including test lengths, sample size, and number of attributes under different Q-matrix structures.
12. **The Use of Predictive Analytics in Test-Optional Admissions of STEM Majors**
   Roseline Telfort; Lin Zeng, Louisiana State University
   This study will present the application of two predictive models, logistic regression and classification & regression tree (C&RT), to determine which variables contribute to admission of prospective STEM majors at a test-optional college/university and which model has the highest classification accuracy of predicting admissions.
13. **Type I Error and Power of Differential Item Functioning Methods: Scoping Review**
   Mubarak Olumide Mojoyinola, The University of Iowa; Brandon LeBeau, University of Iowa
14. **Validity of Students' Evaluation of Teaching in Ghana Using Many-Facet Rasch Model**
   Frank Quansah, University of Cape Coast
   This research presents an understanding of the dependability of students' responses to the evaluation of courses and teaching, which is a usual practice in universities globally. This presentation attempts to unravel the role of rater variability and item characteristics in the ratings of teaching quality in higher education institutions.
15. **Automated Feedback on Interest Development: Mediated by Student Perception of Feedback Usefulness**
   Jan Luca Bahr; Lars Höft, Leibniz Institute for Science and Mathematics Education; Thorben Jansen, Leibniz Institute for Science and Mathematics Education
   In an experimental study (N = 463) within a digital learning environment we compared the impact of automated feedback messages on interest development via student perception of feedback usefulness. Data was analyzed using a latent change scoring model. Results revealed that usefulness mediated the effect from feedback on interest development.

082. Innovating Assessments: Towards Next Generation Assessments of 21st Century Skills

Coordinated Paper Session

2:50 to 4:20 pm

Marriott: Floor 5th - Chicago Ballroom B/C

The assessment of complex constructs (so-called '21st century skills') requires conceptual advances to be combined with technological and methodological advances. In other words, it requires a new generation of assessments capable of measuring complex constructs in innovative ways. The presentations in this Coordinated Paper Session, authored by senior experts in the field of educational measurement, focus on the theme of "Next Generation Assessments of 21st Century Skills". The presentations address key questions including: (1) Why do we need a new generation of assessments? (2) What can we learn from the learning sciences that can help us to better define complex competences for more valid measurement? (3) How can we think about designing assessments of 21st century skills in disciplinary contexts? (4) How can we exploit the affordances of technology when designing innovative assessments? And finally, (5) how we can validate interpretation and uses of assessment results for individuals from different sociocultural contexts? The presentations in this Coordinated Paper Session aim to bring a practical and applied perspective to innovations in assessment design in the context of measuring complex constructs.

Session Organizer:
Natalie Foster

Participants:

From Measuring What Is Easy, To Measuring What Matters
James Pellegrino, University of Illinois at Chicago
Next Generation of Large-Scale Assessments of 21st Century Skills
Mario Piacentini, OECD; Natalie Foster
Complex Problem Solving in STEM: Using Decision-Making as a Guiding Framework
Argenta Price, Stanford University; Carl Wieman, Stanford University
Technology-Enhanced Assessment: The Toolkit for Assessment Designers
Xiangen Hu, University of Memphis; John Sabatini, Institute for Intelligent Systems, University of Memphis
Cross-Cultural Validity and Comparability in Innovative Assessments of Complex Constructs
Kadriye Ercikan, Educational Testing Service; Han-Hui Por, ETS; Hongwen Guo, Educational Testing Service

Discussant:
Jack Buckley, **Roblox Corp**

083. Practical Applications of NLP and Text Mining Techniques for Test Development Tasks

Coordinated Paper Session
2:50 to 4:20 pm
Marriott: Floor 5th - Chicago Ballroom D

With the exponential growth of NLP (Natural Language Processing) and text mining techniques, many industries are experiencing a paradigm shift in the working process, and the measurement field is no exception. Textual information is a critical component that can be used throughout the entire lifecycle of a psychometrically sound assessment, from Job Task Analysis (JTA) to item analysis. For example, text-based information can be used to understand a job role, identify enemy items, flag test content for quality issues, and gain understanding of examinee response behaviors. In the past, these text-related tasks were completed manually by test developers and subject matter experts, if performed at all. Test developers could not utilize their data to its full potential and optimize exam quality due to insufficient analytic techniques and limited computing power. Now that we're fully entrenched in the "Big Data" era, qualitative data can be efficiently analyzed by advanced text mining and NLP techniques. In this session, five practitioners will present how they leverage text mining and NLP to tackle traditionally time-consuming and daunting tasks. Presenters will discuss obstacles encountered in applying these techniques and what was learned during the journey.

Session Organizer:
Huijuan Meng, **Amazon Web Services (AWS)**

Chair:
Anjali Weber, **Amazon Web Services (AWS)**

Participants:

Using NLP to Inform and Enhance Construct Definition and Validation
Vinita Talreja, AWS; Matthew Burke, AWS
Automatic Enemy Identification—Are We There Yet?
Huijuan Meng, Amazon Web Services (AWS)
Analyzing Examinee Comments to Ensure Quality Control in Exam Content
Ye Ma, AWS
A Deep Dive into Process Data for Lab-Based Items
Jennifer Davis, Amazon Web Services (AWS)

Discussant:
Kirk Becker, **Pearson**

084. Psychometric Implications of Item Exposure in Standardized Testing: Investigative Procedures and Impact

Coordinated Paper Session
2:50 to 4:20 pm
Marriott: Floor 5th - Chicago Ballroom E

Owners of a test preparation company in Florida were sentenced to federal prison following a criminal prosecution stemming from the theft of proprietary content from the State of Florida's Teacher Certification Examinations (FTCE) Program. This theft presented a stark threat to the validity of the test scores and the Florida Department of Education took a series of steps to mitigate and react to the potential impacts of this threat on the quality of the teacher candidate pool, as well as to reduce the likelihood of a future security breach. This series of presentations will explore the facts and circumstances of the case, discuss the forensic and other investigative tools used to detect and document the theft, and explore the measurement considerations regarding the impact associated with the exposure of content. The Presenters will also provide an overview of practices that have been put into place to thwart future security breaches. The audience will learn the complex and interdependent nature of security investigations and the severity of the validity threat that such actions have on assessment results, and a governmental agency's duty to respond accordingly.

Session Organizer:
Amy Elizabeth Schmidt, **Pearson**

Participants:
Policy Implications
Phil Canto
Item exposure Impact
Suleyman Olgar, Florida Department of Education
Investigative Steps
Amy Schmidt, Pearson
Preventive Measures
Sarah Underwood, Florida Department of Education

Discussant:
Jon S Twing, **Pearson**

085. Research Blitz: Automated Scoring

Research Blitz Session
2:50 to 4:20 pm
Marriott: Floor 5th - Chicago Ballroom F

Chair:
Janine Jackson, **Morgan State University**

Participants:

Automated Scoring of Argumentation-focused Teaching Transcripts: Added Value of Human Annotations
Duy N. Pham, Educational Testing Service; Viet Lai, University of Oregon; Jamie Mikeska, ETS; Jonathan Steinberg, Test; Heather Howell, ETS; Thien Nguyen, University of Oregon

This study used human annotations and natural language processing (NLP) features to predict human scores from argumentation-focused teaching transcripts. Results indicated that adding human annotated argumentation features to NLP features increased the Quadratic Weighted Kappa (QWK) from .44 to .54, which was close to the human score QWK of .56.

Analysis of Subtrait-Level Marking to Support Automated Scoring of Constructed-Response Math Items
Luis Alejandro Andrade-Lotero, Pearson; Scott Hellman, Pearson; Kyle Habermehl, Pearson

Constructed-response math items have brought additional challenges for automated scoring systems. We present an automated scoring approach that involves the explicit scoring of each subtrait, which are binary subcomponents of the overall rubric. We show that this approach achieves better accuracy in predicting human scores and requires fewer training samples.

Comparison of Human and Computer-Assisted Scoring of Short-Answer Questions
Janet Mee, NBME; Polina Harik, National Board of Medical Examiners; Brian Clauser, National Board of Medical Examiners

This study compares human and computer-based scoring of short-answer questions (SAQs) on a standardized medical education exam. Subject matter experts scored approximately 500 responses for each of 53 items. On average, the results from the system agreed with the subject matter experts 93% of the time.

An Exploration of Automated Scoring of Short Answer Questions in Medical Examinations
Xia Mao, NBOME; Vladimir Kuzinets, NBOME; Mingye Zhao, National Board of Osteopathic Medical Examiners

The study explores the application of automated scoring for short answer questions in high-stakes medical examinations. It analyzes empirical data from the recent administration of a medical licensure examination that newly employed automated scoring. Opportunities and challenges with this application in the context of high-stakes medical examinations are discussed.

An Examination of Automated Essay Scoring DIF on NAEP Reading Items
Mark David Shermis, Performance Assessment Analytics, LLC

This study looked at possible DIF for score predictions among the top three competitors from a recent NAEP-hosted AES scoring competition. DIF was examined for six different demographic variables using three approaches—SMDs, logistic regression, and IRT. The results showed few differences with the patterns displayed by human raters.

Improving Automated Scoring of Prosody Using Deep Learning Algorithm
Kuo Wang, Southern Methodist University; Xin Qiao, Southern Methodist University; George Sammit, Southern Methodist University; Eric C. Larson, Southern Methodist University; Joseph F. T. Nese, University of Oregon; Akhito Kamata, Southern Methodist University

This study proposes and evaluates an improved approach for automated scoring for prosody in oral reading fluency assessment. This research explored different model structures by using deep learning with prosodic and frequency-related acoustic features to produce better performance especially in cross-domain validation, where phrases are assumed to be unknown.

Detailed Prosody Report for Oral Reading Fluency Assessment via Machine Learning
Xin Qiao, Southern Methodist University; Akihito Kamata, Southern Methodist University; Eric C. Larson, Southern Methodist University; Kuo Wang, Southern Methodist University; Sarunya Somsong, Srinakharinwirot University & Southern Methodist University

This study conducted cluster analysis on prosodic features to provide diagnostic information on students' prosodic aspect of oral reading fluency. The results suggested that students with the same rating category exhibited different acoustic characteristics in their reading. It provided further diagnostic descriptions on students' oral reading behaviors.

Automated Speech Scoring using Deep Neural Network Transformer Models
Susan Lottridge, Cambium Assessment, Inc; Christopher Ormerod, Cambium Assessment; Amir Jafari, Cambium Assessment

Most automated speech engines rely on explicitly defined or algorithmic features to produce both the transcription (i.e., conversion of speech to text) and features used to predict scores. This study extends the current research by illustrating the performance of multi-layer neural networks in both transcription and scoring.

086. Research Blitz: Impact of COVID-19

Research Blitz Session
2:50 to 4:20 pm
Marriott: Floor 5th - Chicago Ballroom G/H

Chair:
Kelley Wheeler, **ACS Ventures, LLC**

Participants:

Admission Exam during the Pandemic: Pre-post Comparison in a Mexican University
Melchor Sanchez Mendiola, Universidad Nacional Autonoma de Mexico (UNAM); Manuel García-Minjares, Universidad Nacional Autónoma de México (UNAM); Enrique Buzo-Casanova, Universidad Nacional Autónoma de México (UNAM); Adrián Martínez-González, Universidad Nacional Autónoma de México (UNAM)

Scores in the admission exam to a Mexican university before and during the pandemic were compared. Results of 447,352 students were analyzed. Percent-correct scores were higher in the 2020 cohort (2019=45.8%, 2020=48.1%, 2021=46.7%) (all comparisons $p<0.0001$). Learning loss was not found in students that finished high school in pandemic conditions.

COVID-19 Impact on Racial/Ethnic Group Professional Exam Performance
Ting Wang, American Board of Family Medicine; Thomas O'Neill, ABFM

This study examined the impact of COVID-19 on exam performance among one medical specialty's residents across different race/ethnicity groups. We found the pre-COVID initial score gap and score trajectory maintained during pandemic and there was no differential impact on exam performance across race/ethnicity groups.

Pláticas with Peruvian Teachers about Assessment Knowledge and Practices during Pandemic Times
Maria Vasquez-Colina, Florida Atlantic University

This qualitative study discusses how Peruvian teachers navigate classroom assessment, interactions, and challenges as part of their practices and preparation before and during the COVID-19 pandemic. Results yielded eight themes (classroom assessment, cultural themes, in-service teacher practices, pre-service teacher preparation, school resources, social-emotional support, teacher empowerment and technology resources).

COVID-19 Impact on Physician Assistant National Certifying Examination (PANCE) Performance
Joshua Goodman, NCCPA; Andrew Dallas, National Commission on Certification of Physician Assistants; Andrzej Kozikowski, NCCPA

The COVID-19 pandemic impacted physician assistant programs, and many testing centers were closed or limited the number of examinees to maintain social distancing. This study's primary objective was to determine whether the pandemic impacted first-time examinees' Physician Assistant National Certifying Examination (PANCE) scores and passing rates.

Two Years After COVID-19: Has Student Academic Performance Rebounded?
Aurore Yang Phenow, Data Recognition Corporation

Large-scale state assessments showed decreases in student performance due to COVID-19. 2022 assessment data was analyzed using multilevel modeling to compare student performance two years post-pandemic to pre-pandemic performance to answer the questions, "has student academic performance rebounded back to pre-pandemic level?" and "which student demographics are rebounding?"

Considerations Related to Updating Interim Assessment Norms Following COVID-19
Adam E Wyse, Renaissance Learning

The COVID-19 pandemic has had far reaching impacts on student learning. This study illustrates how updating interim assessment norms using COVID impacted data may influence score interpretations, response-to-intervention (RTI) classifications, and instructional skill recommendations.

Impact of the Covid-19 Pandemic and the Instruction Mode on Student Performance
Joanna Tomkowicz, Data Recognition Corporation; Dong-In Kim, Data Recognition Corporation; Vince Struthers, Data Recognition Corporation; Ping Wan, Data Recognition Corporation

This study investigates an impact of the instruction mode students were exposed to during the Covid-19 pandemic on student performance on the large-scale state assessments in Spring 2021 and whether this impact extended to Spring 2022. Impact of the instruction mode on subgroup performance is also discussed.

087. Differential Item Functioning Detection and More

Paper Session
2:50 to 4:20 pm
Marriott: Floor 5th - Denver/Houston

Chair:
Guangming Ling, **Educational Testing Service**

Participants:

Evaluating a Method for Predicting Language DIF in a High-Stakes, Cross-Language Assessment
Joshua McGrane, University of Oxford; Heather Leigh Kayton, University of Oxford
An exploratory model was built using Random Forest Regression to investigate the extent to which differences in linguistic features of items predict Differential Item Functioning (DIF) across language versions of a high-stakes international assessment. The model showed that textual complexity features explained between 11% and 13% of the DIF variance.

Establishing Criteria for Item Flagging in Diagnostic Classification Models
Selay Zor, University of Georgia; Matthew James Madison, University of Georgia; Laine Bradshaw, Pearson
It is necessary to establish criteria for identifying the degrees to which differential item functioning (DIF) is practically significant. We focus on establishing thresholds based on the degree to which DIF items impact the classification accuracy in diagnostic classification models (DCMs). Based on findings, we propose levels for item flagging.

Using Machine Learning to Identify Causes of Differential Item Functioning
Jeffrey Hoover, University of Kansas; William Jacob Thompson, University of Kansas; Bruce Frey, University of Kansas
We conducted two simulation studies to evaluate a machine learning framework for identifying factors contributing to differential item functioning. We used classification accuracy and area under the curve to evaluate the performance of four machine learning models across a variety of model estimation conditions.

A Framework for Evaluating Fairness in Algorithmic Decision Making: Differential Algorithmic Functioning
Youmi Suk, Teachers College Columbia University; Kyung (Chris) T. Han, Graduate Management Admission
This study proposes a new framework called differential algorithmic functioning (DAF) for algorithmic fairness based on differential item functioning. DAF is defined with a decision variable, a “fair” variable, and a protected variable. We also define two types of DAF—uniform DAF and non-uniform DAF—, and we provide three detection methods.

Comparison of DIF Effect Size Indices in IRT Models
Lavanya Shravan Kumar, University of South Florida; Naidan Tu, University of South Florida; Christopher Nye, Michigan State University; Sean Joo, University of Kansas; Stephen Stark, University of South Florida
To examine the practical importance of differential item functioning (DIF), it is desirable to report effect sizes along with statistical significance. This research explored several DIF effect size indices under a range of conditions using four IRT models– 2PLM, SGRM, GGUM, and MUPP.

Discussant:
Daniel Bolt, **University of Wisconsin - Madison**

088. Issues in the Use of Anonymous Population Data to Infer Learning from Gameplay

Coordinated Paper Session
2:50 to 4:20 pm
Marriott: Floor 5th - Los Angeles/Miami

In this session, we continue our exploration of a novel approach to ascribing content-referenced meaning to students’ scores using Curriculum When gameplay data from millions of children playing learning games are available, but with only a unique anonymous identifier associated with the player, to what extent can such data be used to infer children’s learning from the games? This symposium describes various strategies that can be used to maximize the information gleaned from the data, from technology infrastructure, learning engineering, and combining well-controlled studies with population data. Validity issues, research issues, measurement issues, and practical issues are raised and addressed.

Session Organizer:
Greg Chung

Chair:
Elizabeth Redman, **UCLA CRESST**

Participants:

Validity Issues and Evidentiary Requirements for Game-Based Assessments
Eva L. Baker, UCLA

How Data-Driven Learning Intelligence Powers Large Scale Children's Media Products
Kelly Corrado, PBS KIDS; Cosimo Felline, PBS KIDS; Jessica Younger, PBS KIDS; Jeremy Roberts, PBS KIDS Digital

The Use of Population Data to Measure Learning in Games
Greg Chung; Kilchan Choi, CRESST/UCLA; Elizabeth Redman, UCLA CRESST; Tianying Feng, UCLA; Charles Parks, CRESST/UCLA

Psychometric Models for Measuring Children's Latent Traits Using Gameplay Data
Kilchan Choi, CRESST/UCLA; Tianying Feng, UCLA; Charles Parks, CRESST/UCLA; Greg Chung; Elizabeth Redman, UCLA CRESST

Discussant:
Harold O'Neil

089. Clustered eBoard Session 2

2:50 to 4:20 pm
Marriott: Floor 7th - Salon I

089-1. Clustered eBoard - Creating Short Forms
Electronic Board Session

Participants:

Creating Short Forms of Psychological Scales: Semantic Similarity Matters
Sevilay Kilmen, Bolu Abant Izzet Baysal University; Okan Bulut, University of Alberta; Jinnie Shin

In this study, we compute semantic similarity among items based on BERT embeddings and then use this information when building short forms of a psychological scale. Our findings show that considering the semantic similarity of the items could help researchers build short forms with high measurement accuracy and content diversi**ty.**

How Many Items We Need for 90% Screening Accuracy with Machine Learning?
Yiling Cheng, Kaohsiung Medical University

The effectiveness of Machine Learning methods: logistic regression and classification tree on shortening scales and on predicting the diagnostic outcomes were examined. Using real datasets with AQ10 assessment on subjects with Autism and the control group, the result showed that with five items the short form achieved 90% classification accuracy.

089-2. Clustered eBoard - DIF 1
Electronic Board Session

Participants:

Applying Regularized Estimation to Solve Model Identification Problem in DIF Assessment
Hsiu-Yi Chao, National Taiwan Ocean University; Jyun-Hong Chen, National Cheng Kung University

To solve the model identification problem in DIF assessment, this study applies the lasso estimator for identifying DIF-free items to build a common metric. According to the simulation study, the results indicated that the lasso method can well control Type I error rates with acceptable power under most conditions.

A General Item Response Theory Framework for Latent DIF Detection
Gabriel Wallin; Yunxiao Chen, London School of Economics and Political Science; Irini Moustaki, London School of Economics and Political Science

A flexible modeling framework for latent DIF detection is proposed, where a general item response theory model is combined with a latent class model. We consider a multiple latent group setting, and the DIF-free items are identified through a LASSO penalty in the marginal likelihood function of the model.

Identifying Differential Item Functioning in Polytomous Items with Logistic Discriminant Analysis
Zhuoran Wang, National Council of State Boards of Nursing (NCSBN); William J Muntean, National Council of State Boards of Nursing; Joe Betts, NCSBN; Hao Jia, National Council of State Boards of Nursing (NCSBN)

Logistic discriminant analysis (LDA) combined with item difficulty difference was used to identify differential item functioning (DIF) in polytomous items. The LDA-based method has well-controlled type I error rate and high power, even when there is missing scoring category.

089-3. Clustered eBoard - DIF 2 Electronic Board Session

Participants:

The Effect of Matching Score on DIF Analyses for NAEP-like Assessments

Brian Habing, National Institute of Statistical Sciences; Ya Mo, Boise State University

A DIF analysis was conducted as part of an investigation into NAEP writing prompts' effects on English language learners. Very different outcomes were found when matching based on plausible values and the posterior mean, both using auxiliary information, and the response-based IRT estimates. This is investigated through a simulation study.

Comparison of Matching Variables for Mantel-Haenszel Statistics with Multistage Testing Data

Ru Lu, Educational Testing Service; Paul Adrian Jewsbury, Educational Testing Service; Hongwen Guo, Educational Testing Service

This study investigates the choices of matching variables for the Mantel-Haenszel statistics on the accuracy of DIF detection with multistage testing data. The matching variables include sum scores, theta estimates, and reported scale scores. The results are visualized with a tree diagram and compared with the IRT-based true DIF measure.

When is Subgroup Sample Size Too Small for Meaningful DIF Analysis?

Seohee Park, American Board of Internal Medicine; Yang Zhao, American Board of Internal Medicine

For small sample size scenarios, DIF analysis has limitations due to low power and high positive rates. Using simulated data based on true item parameters from a medical certification exam, this study determines a sample size threshold for meaningful DIF detection for multiple subgroups with small sample sizes.

089-4. Clustered eBoard - DIF 3 Electronic Board Session

Participants:

Intersectional Approach to Differential Item Functioning: Comparing Type I Error Rates

Michael Russell, Boston College; Erin Winters

Intersectionality theory was recently applied to DIF analyses and resulted in a notable increase in flagging of items. Some readers question whether this increased Type I error contributes to this finding. This paper presents results from a simulation study comparing Type I error under the intersectional and the traditional approach.

Lord's Wald χ^2 Test for Differential Item Functioning Evaluation with Multilevel Data

Sijia Huang, Indiana University Bloomington; Dubravka Svetina Valdivia, Indiana University

In this study, a Lord's Wald χ^2 test-based two-stage differential item functioning (DIF) detection procedure for multilevel data was introduced. Metropolis-Hastings Robbins-Monro (MH-RM) algorithm was applied to estimate multilevel IRT models and compute test statistics. The proposed approach showed great power and well.

Intersectional DIF Analyses for a Graduate Student Program Satisfaction Measure

Alejandra Miranda; Michael C. Rodriguez, University of Minnesota

Program satisfaction is important for postsecondary students; therefore, developing psychometrically robust measures is relevant. Using a dataset of 13,000 graduate students, DIF analyses by graduate level and international status were performed and compared to DIF analysis crossing level and international status, as finer groups might lead to meaningful results.

089-5. Clustered eBoard - DIF 4 Electronic Board Session

Participants:

DEMQOL-CH Differential Item Functioning: Do Proxy Assessors' Language and Ethnicity Matter?

Sevilay Kilmen; Sube Banerjee, University of Plymouth; Rashmi Devkota, University of Alberta; Malcolm B. Doupe, University of Manitoba; Emily Dymchuk, University of Alberta; Yinfei Duan, University of Alberta; Carole A. Estabrooks, University of Alberta; Janice Keefe, Mount Saint Vincent University; Jenny Lam, University of Alberta; Hannah O'Rourke, University of Alberta; Seyedehtanaz Saeidzadeh, University of Alberta; Shovana Shrestha, University of Alberta; Yuting Song, University of Alberta; Matthias Hoben, York University

This study examined measurement invariance of the DEMQOL-CH and differential item functioning (DIF) of its items based on proxy assessors' language and ethnicity, and clients' ethnicity. While client's ethnicity did not influence proxy assessor' responses, DIF was found for a small number of items based on assessors' language and ethnicity.

Explanatory Modeling of Language DIF Among Multilingual Learners Using the IRT-C Model

Kevin Krost, Virginia Polytechnic Institute and State University

Differential item functioning (DIF) was evaluated between English- and Spanish-speaking students on released science items from the 2011 TIMSS using the IRT-C model. Several items exhibited DIF, and covariates were modeled to explain DIF. Last, item content features were evaluated to explain any remaining DIF after modeling the covariates.

Extended Test Time Among English Learners: Does Use Correspond to Score Comparability?

Sara Witmer, Michigan State University

Empirical work is needed to determine if extended test time results in score comparability for English learners (ELs). Prior work has examined ELs eligible to use extended time, but not specifically those who use it. NAEP process data will be used to explore results for ELs who use extended time.

089-6. Clustered eBoard - DIF 5
Electronic Board Session

Participants:

Exploring DIF at the Item and Component Levels Using the LLTM

Qingzhou Shi, University of Alabama; Stefanie A. Wind, University of Alabama; Joni Lakin, University of Alabama

This study explores the detection of differential item functioning (DIF) and differential item component functioning (DCF) within a Linear Logistic Test Model (LLTM) context. We conducted simulations with a combination of DIF/DCF proportions and sample size conditions, building from a previous LLTM study of a spatial reasoning assessment.

Performance of Mantel-Haenszel and CATSIB DIF Procedures in Computerized Multistage Testing Environments Christiana

Aikenosi Akande, National Commission on Certification of Physician Assistants; Anne Corinne Huggins-Manley, University of Florida; M. David Miller, University of Florida

This simulation study investigates the performance of Mantel-Haenszel and CATSIB DIF procedures in computerized multistage testing environments. Three factors were manipulated – sample size, impact, and DIF magnitude. Results reveal that MH was conditionally better in terms of Type-I error rates and Power. Results guide best-practices and add to extant literature.

Evaluating Rasch Tree Purification to Improve DIF Detection in Unbalanced Item Conditions

Nana Amma Asamoah, University of Arkansas; Ronna Turner, University of Arkansas; Wen-Juo Lo, Unversity of Arkansas; Kristen Jozkowski, Indiana University; Brandon Crawford, Indiana University

Prior research has shown that in the case of completely unbalanced DIF, the Rasch tree method of DIF detection produces a low proportion of solutions that only include true DIF detection. Iterative purification procedures are applied to determine if more accurate results can be produced.

089-7. Clustered eBoard - English Learners
Electronic Board Session

Participants:

Examining the Use of Text-to-Speech on the Grade 8 NAEP Mathematics Assessment

Yi-Chen Wu, University of Minnesota/NCEO; Martha Thurlow, National Center on Educational

This study examined process data from the National Assessment of Educational Progress (NAEP) grade 8 mathematics assessment to explore the use of text-to-speech (TTS) by four student groups: non-English learners (ELs) without disabilities, non-ELs with disabilities, ELs without disabilities, and ELs with disabilities.

Evaluating the Adequacy of English Learner Reclassification Using a Regression Discontinuity Design

Hanwook Yoo, Educational Testing Service; Mikyung Kim Wolf, Educational Testing Service; Laura D Ballard, Educational Testing Service

This study examined the adequacy of the English learner (EL) reclassification threshold with one state's longitudinal data. Using a regression discontinuity design (RDD), we investigated the reclassification effect of the third graders on their academic achievements. We also evaluated the impact of EL misclassification on the RDD analysis.

089-8. Clustered eBoard - Eye Tracking
Electronic Board Session

Participants:

Using Eyetracking to Address Reading Rates and Speededness on High Stakes Assessments

Ann Arthur, ACT; Jay Thomas, ACT, Inc.

We use eyetracking data to examine the reading rates in Standard Words per minute for students reading and answering questions on high stakes reading and science assessments. Differences in reading rates for passages, graphics, and item stems may support claims about which of Carver's reading gears (1992) test-takers utilize as well as identifying potential speededness issues.

089-9. Clustered eBoard - Instrument Development
Electronic Board Session

Participants:

Measuring Undergraduate Student College Adjustment: Instrument Development

Arlyn Y Moreno Luna, University of California, Berkeley

This instrument captures students' experiences at an elite public institution from campus identity to campus alienation. First, the instrument aims to measure how students adjust to campus in four different components. Second, this instrument hopes to measure how the adjustment levels differ by time spent at the university.

Using a Cultural Equivalency Methodology to Develop an Attitudinal Bilingual Instrument

Diley Hernandez, Georgia Tech; Jayma Koval, Georgia Institute of Technology; Tom McKlin, The Findings Group; Pascua Padro Collazo, University of Puerto Rico; Rafael Arce Nazario, University of Puerto Rico- Rio Piedras; Isaris Quinones Perez, University of Puerto Rico- Rio Piedras; Joseph Carroll Miranda, University of Puerto Rico- Rio Piedras; Taneisha Lee Brown, The Findings Group; Analía Rao, University of California-Irvine; Douglas Edwards, Georgia Institute of Technology; Jason Freeman, Georgia Institute of Technology

Our paper provides an overview of the development of a bilingual (English-Spanish) instrument using a cultural equivalency (CE) assessment development process. The instrument aims to measure attitudinal constructs related to education and motivation and was piloted in a project aimed at broadening participation of K-12 Latinx students in computer science.

089-10. Clustered eBoard - IRT Estimation and Software Electronic Board Session

Participants:

Searching for Interchangeable Parameter Solutions to Increase Equating Stability

Jordan Yee Prendez, American Board of Internal Medicine; Matthew Swain, American Board of Internal Medicine

Limitations in IRT software can occasionally lead to extreme parameters during item calibration. These items can cause problems when they occur in the anchor set. We propose searching the likelihood surface for less extreme solutions that have equivalent or better likelihood values that might improve equating.

Exploring ICL as an Alternative to BILOG-MG on the MCAT® Exam

Ying Jin, Association of American Medical Colleges; Marc Kroopnick, Association of American Medical Colleges; Hyung Jin Kim, University of Iowa; Won-Chan Lee, University of Iowa; Robert Brennan, University of Iowa

The study explored use of ICL as an alternative to BILOG-MG on the MCAT® Exam for operational purposes. The findings suggest that ICL performs quite well in reproducing item parameter estimates from BILOG-MG and accordingly in recovering section and total scale scores for examinees based on BILOG-MG.

089-11. Clustered eBoard - Innovative Bayesian Applications Electronic Board Session

Participants:

Bayesian Regularization in Multiple Indicators Multiple Causes Models

Lijin Zhang, Stanford University; Xinya Liang, Department of Rehabilitation, Human Resources and Communication Disorders, University of Arkansas

To improve the model generalizability, we proposed to use Bayesian regularization to investigate the impact of covariates in multiple-indicators multiple-causes models. Simulation and empirical studies were conducted to evaluate the performance of five Bayesian regularization methods (ridge, lasso, adaptive lasso, spike-and-slab, and horseshoe) in variable selection, parameter estimation, and prediction.

Bayesian Two-Step Factor Score Path Analysis: An Opportunity to Maximize Efficiency?

Linda Galib, Loyola University Chicago; Ken Fujimoto, Loyola University Chicago; Kate Phillippo, Loyola University Chicago; Elizabeth Levine Brown, George Mason University; Naomi Brown, George Mason University

Studies investigating two-step factor score path analysis (FSPA) as an alternative to structural equation modeling (SEM) often use frequentist methods, straightforward measurement and structural models, and simulated data. In applied situations, however, models are often complex. This study addresses more realistic situations by comparing results from Bayesian SEM and FSPA.

Predictive Performance of Bayesian Hierarchical Stacking in Large Scale-Assessments

Mingya Huang; David Kaplan, University of Wisconsin - Madison

This project addresses issues of model uncertainty in the secondary analyses of large-scale educational assessments. We compare Bayesian stacking with a newer extension referred to as Bayesian hierarchical stacking in a simulation study and case study based on PISA. The results indicate Bayesian hierarchical stacking not only obtains better predictive performance but also provides more flexibility in estimating hierarchical models.

089-12. Clustered eBoard - IRT Model
Electronic Board Session

Participants:

Explanatory Item Response Models with Factor Smooth Functions

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

This paper presents explanatory item response models to model nonlinear interaction effects between continuous and binary covariates in person-by-item response data. Factor smooth functions were used to model the interactions. The models were illustrated to understand how highlighting behaviors on text are related to reading comprehension.

Exploratory Measurement Modeling with Lasso: The Role of Measurement Quality

Youngwon Kim, University of Washington; Elizabeth A. Sanders, University of Washington, Seattle

Exploratory approaches for measurement modeling may be useful with large-scale survey and assessment data for which researchers have little theory to guide model selection. The present study used a Monte Carlo simulation to investigate the accuracy of lasso algorithms in fitting a 3-factor model with three levels of measurement quality.

Utility of Models of Distractor Choice in Large-Scale STEM Assessments

Jing Ma, The University of Iowa; Xi Wang; Stephen Dunbar, University of Iowa; Catherine Welch, University of Iowa

This study examines the extent to which IRT models of distractor choice can be leveraged to improve the measurement in large-scale assessment. Fitting these models to data from a recent statewide assessment program showed both diagnostic information and feedback to developers is provided by systematic use of the models examined.

089-13. Clustered eBoard - Random Guessing
Electronic Board Session

Participants:

The Impact of Rapid Guessing on Model Fit and Factor-Analytic Reliability

Alfonso Martinez, University of Iowa; Joseph A. Rios, University of Minnesota

The present study explored differential model fit and factor-analytic reliability (MF&R) in the presence of rapid guessing (RG) across 21 diverse assessment contexts and populations. Specifically, we investigate differential MF&R under the SEM framework when (1) RG is/is not accounted for and (2) under various threshold-based RG identification settings.

A Comparison of Rapid Guessing Scoring Approaches – An Applied Analysis

Jiayi Deng, University of Minnesota, Twin Cities; Joseph A. Rios, University of Minnesota; Alfonso Martinez, University of Iowa

This study aimed to compare four scoring approaches for handling rapid guessing (RG) across five unique assessment contexts. The results suggested that a unidimensional approach that treats RG as missing provides the best model fit; however, a strong association of parameter estimates across all scoring approaches was observed.

089-14. Clustered eBoard - Response Time
Electronic Board Session

Participants:

Comparing Speed-Accuracy Scoring Algorithms in Executive Functioning Measures: A Large Norming Study

Emily Ho; Yusuke Shono, Claremont Graduate University; Berivan Ece, Northwestern University; Richard Gershon, Northwestern University

In a unique, nationally representative sample of US participants (N = 3838) and across two computerized executive functioning measures, we evaluate and compare several scoring algorithms that integrate speed-accuracy, illustrating that the scoring algorithm that uses accuracy corrected for speed has most suitable psychometric properties and best intra-measure consistency.

Incorporating Response Time Using Drift Diffusion Models in an Online Reading Assessment

Klint Kanopka; Jason Yeatman, Stanford University; Amy Burkhardt, Stanford University

Computerized testing affords response time modeling opportunities. One such model, the drift diffusion model (DDM) (Ratcliff, 1978), targets a specific cognitive decision making process. We estimate individual DDM parameters on one reading task and use a machine learning approach to predict scores on another, potentially linking tapped constructs across tasks

089-15. Clustered eBoard - Sampling Design for Field Testing Items in Adaptive Tests Electronic Board Session

Participants:

Online Calibration Design in Computerized Adaptive Testing

Mingqin Zhang, University of Iowa; Catherine Welch, University of Iowa; Stephen Dunbar, University of Iowa

Online calibration embeds pretesting into operational CAT to collect calibration samples for pretest items. In this study, a simulation study was conducted to evaluate different designs of online calibration sampling methods with regards to pretesting and calibration performance of pretest items in CAT.

Comparison of Two Adaptive Sampling Designs for Calibrating Multistage CAT Pretest Items

Shu Jing Yen, Center for Applied Linguistics; Yage Guo, Center for Applied Linguistics

An essential part of maintaining a multistage adaptive CAT program is to replenish test items through pretesting, however, there is a lack of research in designing pretest and in selecting samples for item calibration. This study introduced and compared two innovative adaptive sampling designs for calibrating multistage CAT pretest items using an online calibration approach through embedding field test items in the operational administration.

089-16. Clustered eBoard - Standard Errors Electronic Board Session

Participants:

IRT Score Error Estimates in Item Modeling- Analytical and Bootstrapping Methods

Kamal Chawla, The College Board; Sunhee Kim, College Board; Judit Antal, College Board

One aspect of item modelling that received far less attention in item response theory is computation of standard errors when there is a variance in item parameters. The study focusses on the calculation and comparison of the mean standard error estimates in item modelling using analytical and bootstrapping methods.

Standard Errors for Gaussian Variational Estimation in Multidimensional Item Response Theory

Jiaying Xiao, University of Washington; Chun Wang, University of Washington; Gongjun Xu, University of Michigan

This study applied the updated supplemented expectation maximization (USEM) and the bootstrap method to the Gaussian Variational Expectation Maximization (GVEM) algorithm for standard error estimates of item parameters in multidimensional item response theory (MIRT) models. Simulation results demonstrated the computational efficiency and estimation precision of three GVEM-based standard error estimators.

089-17. Clustered eBoard - Test Security Electronic Board Session

Participants:

Comparison of CUSUM and Change-Point-Analysis Methods to Detect Item Preknowledge

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

This study compares the performance of two CUSUM-based statistics and two CPA-based statistics in detecting item preknowledge using both response and response time information under various conditions. A simulation study and a real data analysis with a linear test are conducted to exhibit the performances of the statistics.

Detecting Comprised Items Using Parallel Logistic Regression Models

Xuechun Zhou, Ascend Learning; Xin Lucy Liu, Ascend Learning

This study aims to develop an ad-hoc item analysis method to identify comprised items. Parallel logistic regression models are built at content domain level. The comprised items are identified by comparing the statistical associations, explanatory power, and the area under the receiver operating characteristics curve (AUC) between the parallel models.

089-18. Clustered eBoard - Validity Electronic Board Session

Participants:

Fairness and Consequential Validity of the ASVAB-Based Learning Capacity Scores

Yixiao Dong, University of Denver; Daniel McNeish, Arizona State University; Denis Dumas, University of Georgia

The present study examines the fairness and consequential validity of the ASVAB-based learning capacity scores. We found allowing and encouraging examinees to take the ASVAB multiple times and utilizing capacity scores can mutually support fair decisions and improve diversity in military recruitment.

Collecting New Validity Evidence for the IXL Real-Time Diagnostic Math Assessment

Xiaozhu An, IXL Learning; Bozhidar M. Bashkov, IXL Learning; Yao Xiong, Roblox Corporation; Christina Schonberg, IXL Learning

A set of studies collected new validity evidence for the IXL Real-Time Diagnostic math assessment, a widely-used interim PreK-12 assessment. The findings supported its internal structure, established multigroup measurement invariance across key student subgroups, and revealed strong correlations between IXL Diagnostic and Florida Standards Assessment scores overall and by subgroup.

Exploring the Validity Properties of The Illinois Educator Preparator Profiles (IEPP)
Mariana Barragan Torres, IWERC, University of Illinois; Meg Bates, IWERC, University of Illinois; Shereen Oca Beilstein, IWERC, University of Illinois
The Illinois Educator Preparation Profiles (IEPP) was designed as an accountability system for teacher preparation programs. In this paper, we analyzed data from Illinois institutions and programs participating in the 2020 IEPP. Our analysis shows that current IEPP measures need a redesign to accomplish their goals of systematically identifying the adequate preparation of teacher candidates and providing programs with information for consistent improvement. We focus on the implications for validity of the lack of variation across indicators related to measurement, and the consequences for equity of using such indicators.

090. Through-Year Assessment and Growth

Paper Session
4:40 to 6:10 pm
Marriott: Floor 5th - Chicago Ballroom B/C

Chair:
Carina M. McCormick, **University of Nebraska-Lincoln**

Participants:

Expected Classification Accuracy for Categorical Growth Models
Daniel Murphy, WestEd; Matthew Brunetti, WestEd; Quintin Love, New Meridian Corporation; Sarah Quesen, WestEd
Categorical growth models describe student growth in terms of performance level category changes, implying some number of misclassifications. This presentation introduces a new procedure for estimating classification accuracy of categorical growth models based on Rudner's (2001, 2005) classification accuracy method for item response theory (IRT) based assessments.

Evaluating Classification Accuracy and Consistency for Categorical Student Growth
Scott Monroe, UMass Amherst; Brendan Longe, University of Massachusetts Amherst
Several states describe student growth using the change in performance level categories from one grade to the next. This study uses multidimensional IRT to examine the classification accuracy and consistency of this categorical student growth. Also, for a given value table, a method for calculating student-level bias is presented.

What's in a Year: Updated Annual Growth Trends on Vertically Scaled Tests
Sanford Student, University of Colorado Boulder
This paper uses 2018-19 results on vertically scaled US state year-end tests of math and ELA to provide updated standardized estimates of growth, using the existence of SBAC to compare variability in growth trends on the same and different tests. Using the same test reduces variability in trends dramatically.

An Analytic Approach to Through-Year Assessments Designs
Garron Gianopulos, NWEA; Yeow Meng Thum, NWEA
We propose an analytic approach for through-year (TY) assessment systems. Using a single state's set of multiple interim scores and a summative test score, we illustrate how a multivariate multilevel model can improve the quality of scores and growth inferences even if scores are not from the same scale.

Aggregating Scores for Summative Assessments: Comparison of Longitudinally-Weighted IRT Proficiency *Estimators*
Aaron Myers, American Board of Internal Medicine; Whitney Smiley Coggeshall, Educational Testing Service (ETS); Jerome Clauser, American Board of Internal Medicine
Assessment systems consisting of multiple, shorter, formative assessments rather than a single, longer summative assessment have garnered increasing attention. Little attention, however, has been given to how scores are aggregated to produce final summative scores. We propose and evaluate four IRT-based longitudinally-weighted proficiency estimators. Implications of weighted scoring are discussed.

Discussant:
Jeff M. Allen, **ACT, Inc.**

091. Gala NCME Comedy Event - 2023

Coordinated Paper Session
4:40 to 6:10 pm
Marriott: Floor 5th - Chicago Ballroom D

A conference highlight is revived this year with the 2023 edition of the NCME Gala Comedy Event. What do weary NCME members need after a long day of paper presentations like "Improving the Estimation of Blatant Traits by Simulating Hyper-sensitive Nano-parameters Using the TikTok Subroutine in R"? That's right!... a time to decompress at a session featuring stand-up comedy, musical entertainment, and scathing psychometric satire--all provided by your measurement colleagues with no fear of the likely effects on their professional careers. It's the perfect intellectual intermission between visits to the exhibit hall for replacement monitor wipes and the hunt for free hors d'oeuvres and drinks at vendor receptions. Come. Laugh. Leave. It's that simple.

Session Organizer:
Gregory Cizek, **University of North Carolina at Chapel Hill**

Participants:

A Psychometrician Walks into a Dinner Party
Kevin Cappaert and Luciana Cançado, Curriculum Associates
15 Tips for International Students Studying in the United States
Xin Li, ACT, Inc.
Satirizational Descantations (of the Academical Variety)
Bill Wraga, University of Georgia
The Evolution of NCME Conferences: What's In, What's Out
Gregory Cizek, University of North Carolina at Chapel Hill
What is the University of Massachusetts Psychometrics Association?
Vafa Alakbarova, Eduardo Crespo Cruz, Lisa Keller, Ketan, Dukjae Lee, Stephen Sireci, Javier Suárez-Álvarez, Dongwei Wang, University of Massachusetts - Amherst

092. [SIGIMIE Session] Diagnostic Measurement: Operational and Implementational Issues

Coordinated Paper Session
4:40 to 6:10 pm
Marriott: Floor 5th - Chicago Ballroom E

Since the publication of the seminal book Diagnostic Measurement: Theory, Methods, and Applications (Rupp, Templin, Henson, 2010), there has been a wave of research on diagnostic measurement. Much of this research, however, has focused on methodological advancements and applications. Recently, there has been increased implementation of large-scale diagnostic assessments systems in K-12 settings, resulting in a need for more research on issues faced in these settings. This session, coordinated by the Diagnostic Measurement SIGMIE, presents a selection of research studies focused on issues faced in the operation of diagnostic assessments systems and their implementation. More specifically, the studies examine critical issues such as the estimation of diagnostic score reliability, validation of diagnostic results, integration of diagnostic assessment into curricula, and scoring models for diagnosing foundational skills. The session will conclude with a semi-structured question and answer panel to discuss these issues in more depth.

Session Organizer:

Matthew James **Madison, University of Georgia**
Yu Bao, James **Madison University**
Qianqian Pan, **National Institute of Education, Nanyang Technological University**

Participants:

Integrating Diagnostic Assessment into Curriculum: A Theoretical Framework and Teaching Practices
Tingting Fan, Nanjing University; Jieqing Song, Foreign Language Teaching and Research Press
A Simulated Retest Methods for Estimating Classification Reliability
W. Jake Thompson, University of Kansas; Amy Clark, ATLAS: University of Kansas
Generating Trustworthy Measurement Results About Instructionally Relevant Attributes
Daniel Katz, NWEA; Yon Soo Suh, NWEA; Meredith Langi, NWEA; Tyler Matta, NWEA
Effect of Foundational Reading Skills Domain Ordering in a National Diagnostic Reading Assessment
Aimee Boyd, Curriculum Associates; Laurie Davis, Curriculum Associates

093. Assessing Non-cognitive Traits with Multi-dimensional Forced-choice Assessments: Design, Development, and Validation

Coordinated Paper Session

4:40 to 6:10 pm

Marriott: Floor 5th - Chicago Ballroom F

Forced-choice assessments where a respondent chooses among a set of two or more statements the one that most (and/or least) describes him/her, have gained traction for assessing non-cognitive traits/skills because it is believed to be more robust to faking/gaming (Drasgow, et al, 2012; Jackson et al., 2000). This coordinated paper session includes four papers focusing on the design, development, and validation of a multi-dimensional forced-choice (MFC) assessment of interpersonal and intrapersonal skills essential to higher education and career success. The papers speak to the full set of technical considerations/evaluations going into the design and development of a psychometrically sound assessment with this innovative test format: 1) the design of content, item format, and the test, 2) the evaluation and introduction of new estimation/modeling approaches for MFC tests, 3) regression based validity studies to evaluate the relationship of the assessment with other outcomes, and 4) the evaluation of fairness through subgroup comparisons and differential item functioning (DIF) analysis. Information and findings shared will provide practical suggestions and solutions for the use of MFC tests for assessing non-cognitive traits.

Session Organizer:
Xuan (Adele) Tan, **ETS**

Chair:
Ou Lydia Liu, **ETS**

Participants:

Assessing Interpersonal/Intrapersonal Skills for Admissions: Content, Format, and Assembly
Patrick Charles Kyllonen, ETS; Xuan (Adele) Tan, ETS; Daniel Fishtein, Educational Testing Service; Serguei Denissov, ETS; David Schor, ETS; Harrison Kell, ETS; Qi Diao, ETS

In this talk we review, for MFC tests, the processes of (a) selecting the most critical dimensions to assess, (b) evaluating different formats, (c) evaluating susceptibility to faking using different instructions, and (d) assembling statements into blocks to maximize information yet mitigate faking by controlling social desirability within a block.

New Developments/Comparison of Item Response Theory Models for Multi-dimensional Forced-choice Questionnaires
*Jianbin Fu, Educational Testing Service; Xuan (Adele) Tan, ETS; Patrick Charles Kyllonen, ETS; Steven Holtzman, Educational Testing Service; Nimmi Devasia, Educational Testing Servic*e

Several IRT models for MFC questionnaires are compared in terms of model selection, model fit, item fit, reliabilities and correlations of trait score estimates. New direct estimation methods by MML-EM were developed. Based on comparison results, recommendations were made on the models to use for different MFC item formats.

Validating a Noncognitive Assessment Using Multi-dimensional Forced-choice Item Format
David Klieger, ETS; Patrick Charles Kyllonen, ETS; Steven Holtzman, Educational Testing Service; Nimmi Devasia, Educational Testing Service

Regression analyses were conducted to determine the degree to which scores on a multi-dimensional forced-choice (MFC) non-cognitive skills assessment added to the prediction of outcome beyond the prediction given by the traditional predictors (pre-admission grades/test scores). Results provided support for the value of the trait scores obtained from the assessment.

Fairness Evaluation of a Noncognitive Assessment Using Multi-Dimensional Forced-Choice Item Format
Xuan (Adele) Tan, ETS; Jianbin Fu, Educational Testing Service; Patrick Charles Kyllonen, ETS; Steven Holtzman, Educational Testing Service; Nimmi Devasia, Educational Testing Service

This paper 1) extended the logistical regression method for DIF (LR-DIF) to forced-choice tests and 2) examined subgroup mean differences on the assessment across two demographic variables (gender, race) against other standardized achievement tests through independent t-test and effect size measures.

Discussant:
Brent Bridgeman, **ETS**

094. Computer Adaptive Testing: Variations and Impacts

Paper Session
4:40 to 6:10 pm
Marriott: Floor 5th - Chicago Ballroom G/H

Chair:
Cornelis Potgieter, **Texas Christian University**

Participants:

Impact of Shifting Student Performance on Starting Theta and Ability Estimation Accuracy
Jinah Choi, Edmentum, Inc.; Sonya Powers, Edmentum, Inc.

Real and simulated data are used to illustrate impact of using prior achievement as starting values in a computerized adaptive test (CAT) when current achievement has shifted. Impact is evaluated based on score bias, score precision in fixed length CAT, test length in variable length CAT, and item exposure.

The Effect of Theta Initialization Error on Multistage Simulations
Elizabeth Ayers-Wright, Cambium Assessment; Carol Woods, Cambium Assessment; Tao Jiang, Cambium Assessment, Inc; Christina Schneider, Cambium Assessment, Inc.

We investigate if measurement error in the standard deviation of ability estimates should be corrected when the assumed ability distribution for a simulation is from an operational data set in a through-year context. Implications for accurate routing through a multi-stage through-year assessment design are addressed.

Utilizing Response Time for Item Selection in On-the-fly Multistage Adaptive Testing
Xiuxiu Tang; Joshua Goodman, NCCPA; Fen Fan, NBME; Hua-Hua Chang, Purdue University

On-the-fly multistage adaptive testing (OMST) has drawn researchers' interest since it can balance the advantages and limitations of CAT and MST. This study plans to incorporate response time into OMST to select items for a certification exam to improve its measurement efficiency.

Identifying Methods for Multistage Testing (MST) Routing with Mixture Modeling
Ahmed Bediwy, The University of Iowa; Cassondra Griger, University of Iowa; Jonathan Templin, University of Iowa

Misrouting examinees in a Multistage Test (MST) design results in inaccurate ability scoring, especially in high-stake assessments. This study explores differing methods for designing an effective routing stage of a two-stage MST using IRT mixture modeling (MixIRT) analysis on a mathematics data set.

Discussant:
Laurie Davis, **Curriculum Associates**

095. GSIC Standards Study Group: Recommendations from Graduate Students for Its New Version

Paper Session
4:40 to 6:10 pm
Marriott: Floor 5th - Denver/Houston

In September 2021, the Graduate Student Issues Committee (GSIC) convocated a study group for graduate students about the current version of the Standards for Psychological and Educational Testing. The group met weekly during Fall and Spring 2021 and Summer 2022. The group covered each chapter, switching between discussion sessions and Q&A sessions with experts. In this Discussion session at NCME, we will cover the main takeaways of the study group and its implications for the new version of the Standards, which is soon to start its writing process. We will cover five main topics: Fairness, Validity, Psychometrics, and Test Development. We will have the students that hosted the chapters' sessions for each topic. We also included one new section/topic about Machine learning and Artificial Intelligence that, we believe, should have its chapter in a future version of the Standards. Kristen Huff will be the final Discussant of the session.

Session Organizer:
Sergio Araneda, **University of Massachusetts Amherst**

Moderator:
Janine Jackson, **Morgan State University**

Presenters:
Merve Sarac, **UW-Madison**
Magdalen Beiting-Parrish, **CUNY Graduate Center**
Montserrat B Valdivia Medinaceli, **Indiana University Bloomington**
Stacy R Huff

Discussant:
Kristen Huff, **Curriculum Associates**

096. Methodological Advances in Detecting and Accounting for Noneffortful Responding

Coordinated Paper Session

4:40 to 6:10 pm

Marriott: Floor 5th - Los Angeles/Miami

Noneffortful responding (NER) results in responses that are unreflective of what examinees know and can do, thus posing a major threat to the validity of assessment results. Current approaches capitalize on the collateral information in response times, assuming that they signal differences in examinees' response strategies. Although there is a rich body of response-time-based approaches to NER, two major methodological challenges remain: First, there is yet no agreement on optimal NER detection techniques. Second, once identified, researchers need to decide on the adequate treatment of NER. This symposium highlights five methodological advances for response-time-based detection and handling of NER addressing these issues. The first paper introduces a data-driven approach to inform response time thresholds distinguishing NER from solution behavior. The second paper develops a probabilistic filtering procedure that takes the uncertainty in NER identification into account. The third paper presents a simulation study that evaluates different approaches to both rescoring and model-based treatment of flagged responses. The fourth paper investigates the tenability of behavioral assumptions underlying modeling approaches to handling NER in a large corpus of large-scale assessment data sets. The fifth paper explores adjusting for differences in examinees' effort by conditioning item and person parameters on response times.

Session Organizers:

Esther Ulitzsch, **Leibniz Institute for Science and Mathematics Education**
Joseph A. Rios, **University of Minnesota**

Participants:

Detecting Rapid Guessing: In Search of the Optimal Response Time Threshold
Guher Gorgun, University of Alberta; Okan Bulut, University of Alberta; Tarid Wongvorachan, University of Alberta

Researchers often identify rapid guessing by using a fixed normative threshold (e.g., 10% of the average response time) across all items while neglecting the role of item characteristics. In this study, we aim to harness iterative optimization procedures to search for the optimal response time threshold for each item.

A Probabilistic Filtering Approach to Accounting for Noneffortful Responding
Esther Ulitzsch, Leibniz Institute for Science and Mathematics Education; Joseph A. Rios, University of Minnesota; Radhika Kapoor; Klint Kanopka; Ben Domingue, Stanford University

The presented probabilistic filtering procedure to noneffortful responding (NER) allows taking uncertainty of response-time-based NER classification into account. This is achieved by pooling analysis results from multiple plausible representations of filtered data sets. The procedure is outlined and illustrated based on two different approaches to creating filtered data sets.

A Comparison of Response Time Threshold Scoring Procedures for Handling Rapid Guessing
Joseph A. Rios, University of Minnesota; Jiayi Deng, University of Minnesota

The present simulation study compared item and ability parameter recovery for four response time threshold scoring procedures by manipulating sample size, the linear relationship between rapid guessing (RG) propensity and ability, percentage of RG responses, as well as the type and rate of RG misclassifications.

Evaluating Modeling Assumptions Around Rapid Guessing: Results from a Low-Stakes Assessments Corpus
Alfonso Martinez, University of Iowa; Joseph A. Rios, University of Minnesota

Data were collected for 20 low-stakes assessments to investigate: (a) the extent to which rapid guessing (RG) propensity is linearly related to ability and the strength of association; and (b) whether expected response probability predicts RG or if RG is reflective of an idiosyncratic decision for each examinee-by-item interaction.

Sensitivity of PISA Item and Person Parameter Estimates to Differential Response Times
James Soland, University of Virginia; Joseph A. Rios, University of Minnesota; Rujun Xu, University of Virginia

Person and item parameter estimates can be sensitive to how long students spend on a given item. We investigate this sensitivity by conditioning PISA item and person parameters on response times. Results suggest that PISA scores can differ substantively dependent on whether or not parameters are conditioned on response times.

Discussant:

Steven Wise, **NWEA**

097. eBoard Session 2
Electronic Board Session

4:40 to 6:10 pm
Marriott: Floor 7th - Salon I

Participants:

1. **Using Aggregated Residuals as Proxies for Factor Scores: A Cautionary Tale**
Jonathan Weeks, Educational Testing Service
This study uses empirical data from a measure of six foundational reading skills with an established multidimensional structure. Aggregated residuals from a unidimensional IRT model were created as proxies for the factor scores and used in a series of regression models. The results suggest these types of proxies are problematic.

2. **Long-term English Learners' Test Scores and Grades**
Nami Shin, ATLAS, University of Kansas
This study examines how long-term English Learners' (LTELs) scores on two annual assessments, English language proficiency (ELP) and content assessments, are compared to their grades at school. Analyzing longitudinal student-level data from a large urban school district, this study describes LTELs' performance trajectories in the assessments and school grades.

3. **Exploring the Relationship of Students' Perception of Online Testing and Academic Performance.**
Tavia Flowers
A common source of variability is cognition therefore variability among students' perception is expected and meaningful. from a measurement perspective, emphasis is placed on reducing variability to increase the dependability of observed consistency of score. Therefore, this study is proposed to explore this phenomenon in association with online summative assessment.

4. **A Monte Carlo Simulation of Scoring Methods for Multiple Response Items**
Yin Burgess, National Registry of EMTs; Mihaiela Ristei Gugiu, National Registry of Emergency Medical Technicians
Multiple response items have the potential to provide additional information about candidates' performance. Typically, these items are scored through cluster scoring using "all or nothing" or using one credit for each correct option. This Monte Carlo simulation study seeks to compare the two methods and their impact on candidate ability.

5. **Examining Relationship Between Item Writing Productivity and Item Quality**
Aijun Wang, FSBPT; Yu Zhang, Federation of State Boards of Physical Therapy; Lorin Mueller, Federation of State Boards of Physical Therapy
This study explored whether item writers become better as they write more items. We examined the pretest survival rate using a logistic regression model and the relationship between item's characteristics and the number of items that were written by the item writers using multilevel regression models.

6. **Investigating the Effects of Programming Task Characteristics on Student Programming Process**
Mo Zhang, Educational Testing Service; Min Li, University of Washington; Amy Ko, University of Washington; Benjamin Xie, University of Washington; Hongwen Guo, Educational Testing Service; Paul Pham, University of Washington; Jared Lim, University of Washington
Learning to code is becoming a popular subject for learners of all ages. Yet, educators generally agree that computer programming is difficult to teach and assess. This paper will present a small-scale pilot study that aims to address difficulties in assessing computer programming by investigating critical characteristics of programming tasks.

7. **Developing Numeracy Diagnostic Assessment to Support Teaching and Learning**
Girts Burgmanis; Dace Namsone, University of Latvia
This study provides information how to develop numeracy diagnostic assessment using three-dimensional framework grounded in curriculum based on 21st century skills and how test results could be used as communication tool to support teachers to adjust instructional practices, track student progress and identify students for intervention support.

8. **Investigating Predictive Validity of a Medical Licensing Examination Using Multilevel Modeling**
Mohammed Abulela; Irina Grabovsky, National Board of Medical Examiners
We examined the predictive validity evidence of the United States Medical Licensing Examination®-Step 2 Clinical Knowledge Component with patient mortality using multilevel logistic regression. We utilized deidentified data of 150,907 patients nested within 1,744 physicians nested within 170 hospitals. Results revealed that higher scores were associated with lower patient mortality.

9. **Effect of Insufficient Effort Response on Test Validity Using a Multiple-Hurdle Approach**
Yelin Gwak; Youn-Jeng Choi, Ewha Womans University
This study investigates the impact of insufficient effort response on test reliability and validity using a multi-hurdle approach. We used TIMSS 2019 background survey data of eighth graders in the United States. Insufficient effort response was detected and removed by sequentially applying several insufficient effort response detection methods.

10. **Effectiveness of a Virtual Reality Simulation as an Assessment Tool in Online Competency-Based Higher Education**
Sean Gyll, Western Governors University
Virtual reality (VR) assessments represent a much-needed effort to move beyond the shortcomings of today's forms-based measures. Within VR, we assess for competency and problem-solving skills versus the content memorization typically supported by multiple-choice assessments. This paper demonstrates an innovative VR assessment recently deployed in Western Governors University College of Health Professions. It follows Emil, a 32-year-old patient undergoing treatment for Type II Diabetes and highlights several of the design and measurement considerations important in VR assessment. We investigated students' summative assessment scores across a 2D (desktop) and 3D VR (headset) version and how their scores were impacted by motion sickness, cognitive workload, and system usability issues. Several practical implications to aid assessment professionals in developing VR examinations are provided.

11 **How to Execute Routine Psychometric Tasks Automatically on a Schedule**
Hotaka Maeda, Smarter Balanced
Despite not commonly discussed, scheduling programmed tasks to be automatically executed is relatively easy for individual psychometricians with some programming experience. This can save a considerable amount of time for the effort invested. I present some necessary steps and considerations for scheduling tasks, with examples.

12. **Dimensionality of Ngss-Aligned Science Tests**
Sakine Gocer Sahin, New Meridian Corporation; Donna J Butterbaugh, (ISC)2
Construct validity is an important feature of educational instruments. The purpose of this study is to evaluate the dimensionality of New Meridian NGSS-aligned Science tests using data from one state administration at three grade levels. A bifactor model within the multidimensional IRT (MIRT) framework was used to evaluate dimensionality.

13. **Evaluating the Specification of IRT Proficiency Estimators for Long-Term Score Accuracy**
Stella Kim, University of North Carolina at Charlotte; Won-Chan Lee, University of Iowa
This study examines possible specifications of IRT proficiency estimation methods and evaluates them with respect to long-term accuracy of estimates. Four IRT proficiency estimation methods are investigated: a) test characteristic function with number-correct scoring, b) MLE with pattern scoring, c) EAP with number-correct scoring, and d) EAP with pattern scoring.

14. **Examining the Dimensional Structure Between Orthographic Processing and Word Reading: A Meta-Analysis**
Songtao Wang, OISE/University of Toronto; Krystina Raymond, University of Toronto; Zein Abuosbeh, University of Toronto; Diana Burchell, University of Toronto; Becky Chen, University of Toronto
This study examined whether the correlation between orthographic processing and word reading skills is latently dimensional or categorical. We synthesized 57 findings drawn from 16 articles that reported empirical results of Pearson's r. Random-effects models indicated that the relationship is dimensional rather categorical, indicating a small to medium effect size.

15. **Identifying Latent State Transitions with Multigroup Hidden Markov Model on Process Data**
Ni Bei, University of Washington; Qiwei He, Educational Testing Service; Yang Jiang, ETS
This paper proposes a multigroup hidden Markov model (HMM) on sequential process data by employing background covariates (e.g., basic literacy skills) to identify and visualize the latent state transitions by different groups, using action sequences from 1,338 US respondents in a scenario-based interactive problem-solving item in PIAAC.

098. NCME Fitness Run/Walk (Registration Required)

6:00 to 7:00 am
Marriott: Meet in Hotel Lobby

099. Computer Adaptive Testing: Item Pool Development and Calibration

Paper Session
8:00 to 9:30 am
Marriott: Floor 5th - Chicago Ballroom B/C

Chair:
Emily Ho

Participants:

Item Pool Development and Maintenance
Dipendra Subedi, Pearson Assessments; Changjiang Wang, Pearson; David Shin, Pearson

This study presents a multiphase multiyear approach to item pool development and maintenance for successful implementation of CAT. Using methodology currently employed in two operational CAT programs, we present a robust plan on item pool maintenance and demonstrate the item parameter drift (IPD) results from an operational CAT program.

New 1-bit Matrix Completion-Based Methods for CAT Item Bank Calibration
Yawei Shen; Shiyu Wang; Houping Xiao, Georgia State University

To address the challenges of calibrating an item bank with small sample sizes, this study proposed new methods based on 1-bit matrix completion. from a simulation study, the proposed methods led to better item discrimination parameter estimations than the baseline methods and comparable item difficulty parameter estimation.

Counting All Possible Test Forms
Mengyao Cui, Cambium Assessment, Inc.; Widad Abdalla; Frank Rijmen, Cambium Assessment, Inc

Four methods for counting all possible forms given an item bank and a blueprint are proposed. These methods help test developers determine whether an item pool is deep enough to support a desired test design, identify shallow areas of the item pool, and provide guidelines for future item development.

Discussant:
Yanyan Fu, GMAC

100. Better Decisions Through Comprehensive Statistical Model Evaluation

Coordinated Paper Session
8:00 to 9:30 am
Marriott: Floor 5th - Chicago Ballroom D

Educational researchers often rely on statistical models to represent and test their theories about the relationships among the variables associated with the phenomena of interest. Such models are central to scientific analysis, inference, and decision-making, so it is imperative that they have been carefully and thoroughly evaluated prior to dissemination. In this organized paper session, four speakers will discuss and demonstrate several recent advances in statistical model evaluation. Wes Bonifay (University of Missouri) will present a theoretical framework that integrates traditional, Bayesian, and information-theoretic methods and thereby yields a blueprint for comprehensive statistical model evaluation. Sonja Winter (University of Missouri) will demonstrate a previous unexamined contributor to model complexity: the impact of Bayesian prior specification on the model's inherent propensity to fit well to diverse data patterns. Yon Soo Suh (NWEA) will compare full- and limited-information approaches for quantifying parsimony in item response theory models, with an emphasis on investigating the feasibility of the limited-information method. Finally, Li Cai (UCLA) will demonstrate how the computational capability of this limited-information approach facilitates information-theory-based analyses of models intended for many items and/or polytomous response scales. Derek Briggs (University of Colorado Boulder) will be the discussant for this coordinated paper session.

Session Organizer:
Wes Bonifay, University of Missouri

Participants:

Comprehensive Statistical Model Evaluation in the Education Sciences
Wes Bonifay, University of Missouri; Sonja D Winter, University of Missouri; Hanamori Skoblow, University of Missouri
Prior Specification in Bayesian Estimation Affects a Model's Fitting Propensity
Sonja D Winter, University of Missouri; Wes Bonifay, University of Missouri
Model Complexity in Item Response Models: Full- versus Limited-Information Approaches
Yon Soo Suh, UCLA
Using Limited-Information Methods to Assess the Fitting Propensity of Polytomous Item Response Models
Li Cai, UCLA

Discussants:

Derek Christian Briggs, **University of Colorado Boulder**

101. The Future is Now: Game-Changing Innovations in Educational Assessment

Coordinated Paper Session
8:00 to 9:30 am
Marriott: Floor 5th - Chicago Ballroom E

There are exciting innovations in educational assessment that are not yet well known by the measurement community, in part because they are brand new and involve unique collaborations among psychometrics, computer scientists, and social justice researchers. In this session, we highlight and illustrate four examples of some of the most significant advances in educational measurement that are likely to change educational assessments for the foreseeable future. One presentation illustrates how a 21st-century perspective on Lord's concept of randomly parallel tests led to the development of a computerized, real-time item generation system. A second provides examples from computational psychometrics that illustrate next-generation item development and calibration strategies that involve minimal pilot-test demands. The third illustrates a reconceptualization of score reporting that leverages historical item meta-data and student characteristics to provide more information about students' knowledge and skills to help increase student achievement. The fourth introduces the concept of a test assembly system that builds tests for multiple purposes from an assessment task warehouse, with all tests being customizable to the learner. The presentations go beyond proposing new ideas by showing innovations that are already taking place.

Session Organizer:

Stephen Sireci, **University of Massachusetts Amherst**

Chair:

Francis O'Donnell, **National Board of Medical Examiners**

Participants:

Randomly Parallel Tests and SmartItems: A Synergistic Combination of Old and New Innovations
David Foster, Caveon Test Security
Explanatory Item Response Models with BERT Contextual Word Embeddings in Language Assessment
Kevin Yancey, Duolingo; Andrew Runge, Duolingo; Geoff LaFlair, Duolingo; J.R. Lockwood, Duolingo
Predictive Diagnostics for Flexible and Efficient Assessments
Thomas Christie, NWEA; Anna Rafferty; Carson Cook, NWEA; Hayden Johnson, MN
DIRTy CATs and Other DIRTy Assessments: The Adult Skills Assessment Program
Javier Suárez-Álvarez, University of Massachusetts Amherst; Maria Elena Oliveri, University of Nebraska Lincoln; Stephen Sireci, University of Massachusetts Amherst

Discussants:

April Zenisky, **University of Massachusetts Amherst**

102. Integrating Process Data in Psychometric Models

Paper Session
8:00 to 9:30 am
Marriott: Floor 5th - Chicago Ballroom F

Chair:

Hongwen Guo, **Educational Testing Service**

Participants:

Binary Time Series for Eye Fixations: The Quality of Parameter Estimates
Paul De Boeck, OSU; Selena Wang, Yale School of Public Health; Sun-Joo Cho, Peabody College of Vanderbilt
Binary time series are a fine-grained approach for eye fixations in cognitive tasks and choice tasks. Based on simulations we give indications for a sufficient quality (precision, bias) of time series parameter estimates (intercept, trend, autocorrelation). A real data application also gives us indications on the precision of parameter estimates

How Factor Models Recover Response and Response Time Generated from Q-diffusion Models
Chen Tian
Q-diffusion model is a process IRT model that incorporates the underlying cognitive process when respondents are deciding which answer is correct. This study examines the cross fitting of joint factor models and the Q-diffusion Models. Resulted discrepancy shed lights on the conceptual differences in the definition of latent variables.

Development and Calibration of a Rasch Model Integrating Process Data from Topic Modeling
Jiawei Xiong, Pearson; George Engelhard, UGA; Allan Cohen, University of Georgia
This paper proposes a Rasch measurement model that integrates process data extracted from constructed response items into the partial credit model. Parameters were estimated through Hamiltonian Monte Carlo. This model is evaluated using Stan with real data in R. Potential applications of this model are discussed.

Joint Bi-factor Modeling of Item Responses, Response Time, and Answer Changes
Hong Jiao, University of Maryland; Dandan Liao, McKinsey & Company
Joint modeling of item product and process data increases IRT model parameter estimation accuracy and reveals the relationship between latent ability and other process latent traits. This study proposes joint bi-factor modeling of responses, response time, and answer change frequencies to increase the accuracy of latent ability parameter estimation.

Exploring Added-Value of Response Times for Low-stakes Assessments with Nested Logit Models
Cigdem Bulut; Okan Bulut, University of Alberta
This study examines whether the Nested Logit Item Response (NLIRT) model can be used to extract additional information from item response times to enhance the accuracy of low-stakes assessments. The results showed that NLIRT could yield higher measurement accuracy—especially for low-ability students—when compared with logistic IRT models.
.

Discussant:
Leah Feuerstahler, **Fordham University**

103. Advanced Technology Use in TIMSS and PIRLS

Paper Session
8:00 to 9:30 am
Marriott: Floor 5th - Chicago Ballroom G/H

The introduction of new technologies and psychometric modeling approaches, as well as the collection of log and process data in computer-based assessments, triggered a number of exciting innovations and advances in test development, scoring, and data analysis in the Progress in International Reading Literacy Study (PIRLS) and the Trends in International Mathematics and Science Study (TIMSS). The papers in this session illustrate the use of Machine Learning algorithms for automated item generation and automated scoring of constructed image and text responses, as well as the use of log and process data to improve our understanding of response behavior and the estimation of achievement scores in international large-scale assessments.

Session Organizer:
Lale Khorramdel, **Boston College**

Chairs:
Lale Khorramdel, **Boston College**
Matthias von Davier, **Boston College**

Participants:

Automated Reading Passage Generation using OpenAI's GPT-3
Ummugul Bezirhan, Boston College TIMSS & PIRLS International Study Center; Guillermo Ravelo, Boston College; Matthias von Davier, Boston College

Automated Scoring of TIMSS 2019 Graphical Responses using Convolutional Neural Networks
Lillian Tyack, Boston College; Lale Khorramdel, Boston College; Matthias von Davier, Boston College

Automated Scoring of Multilingual Written Responses using Artificial Intelligence
Ji Yoon Jung, Boston College; Lillian Tyack, Boston College; Matthias von Davier, Boston College

Using Process Data to improve the Estimation of Achievement in PIRLS 2021
Dihao Leng, Boston College; Lale Khorramdel, Boston College; Matthias von Davier, Boston College

Process Data for Measurement: A Systematic Literature Review
Ella Anghel; Lale Khorramdel, Boston College; Matthias von Davier, Boston College

Discussant:
David Rutkowski, **Indiana University**

104. Method and Conceptual Development in Test Scaling, Linking, and Equating

Paper Session
8:00 to 9:30 am
Marriott: Floor 5th - Denver/Houston

Chair:
Luciana Cancado, **Curriculum Associates**

Participants:

Using Normalized Theta Score Differences to Evaluate Equating with Item Parameter Drifts
Yong He, Measurement Incorporated; Troy Chen, Measurement Incorporated

This study investigates the magnitudes of normalized theta score differences (Luecht, 2016) to evaluate equating results with item parameter drifts detected by displacements and t-test (Liu & Jurich, 2022) approaches. Simulation study results will be analyzed to inform psychometric practitioners of scale drift and quality of equating results.

The NEAT Equating via Chaining Random Forests: A Machine-learning Method
Yuting Han; Zhehan Jiang; Lingling Xu, Peking University; Jinying Ouyang

We present a machine-learning-based (ML-based) imputation technique called Chaining Random Forests (CRF) to perform equating tasks within the nonequivalent groups with anchor test (NEAT) design. The simulation study suggests that certain CRF-based methods can yield more accurate equated scores than other counterparts in short-length tests with small samples.

Scale Transformation with Variance Stabilization Using Higher-Order Polynomials
Judit Antal, College Board

This paper introduces a scale transformation method that uses reparameterized higher-order polynomial functions to match predened target distributions, which also controls the error variance of the transformed scores. This is achieved by numerical optimization, which is demonstrated through three studies. Supportive evidence is provided for the effectiveness of the proposed procedure.

Population Invariance in Composite-score Equating with the Random Groups Design
Kuo-Feng Chang; Won-Chan Lee, University of Iowa

A growing body of literature has emerged on composite-score equating. However, there is a paucity of literature that investigates the performance of a variety of composite equating procedures with respect to population invariance. This study addresses the complex issues associated with population invariance in the context of composite equating.

Integrating Measurement Error and Equating Error
Won-Chan Lee, University of Iowa; Stella Kim, University of North Carolina at Charlotte

This study proposes a general procedure for quantifying overall error involved in equated test scores under the classical test theory and IRT frameworks. Two types of errors in equated scores (i.e., equating error and measurement error) are integrated in a single error index.

Discussants:
Yue Jia, **Educational Testing Service**

105. Classroom and Instructionally Embedded Assessment

Paper Session
8:00 to 9:30 am
Marriott: Floor 5th - Los Angeles/Miami

Chair:
Guher Gorgun, **University of Alberta**

Participants:

Cultural Validity: Promoting Cultural Responsiveness in Classroom Assessment
David Baidoo-Anu; Lei Liu, Educational Testing Service; Dante Cisterna-Alburquerque, ETS; Yi Song, Educational Testing Service

This study aims to methodically map extant literature to conceptualize cultural validity in classroom assessment. Three themes emerged — conceptualization of cultural validity, promoting cultural validity, and challenges to cultural validity. This review provides a foundation for better supporting teachers to effectively implement and engage in culturally responsive assessment.

Promises and Challenges in Teacher Implementation of Instructionally Embedded Assessments
Amy Clark, ATLAS: University of Kansas; Jennifer Kobrin, ATLAS: University of Kansas; Megan M Mulvihill, University of Kansas; Ashley Hirt, ATLAS at the University of Kansas

Instructionally Embedded (IE) assessments can both provide timely assessment data for instruction and meet summative reporting needs. However, there is limited research describing teacher use of these assessments. We describe promises and potential pitfalls gleaned from teacher focus groups for an operational IE assessment system, including implications for other programs.

Ontology-based Reasoning for Classroom Assessment
Yoav Bergner; Ofer Chen, New York University
The use of ontologies for student assessment by teachers in student-centered learning environments was explored using a focus group design. The work was motivated by Wilson's argument about coherence between instructional values and assessment systems and how a 'community of judgment' needs to place teachers in a central position.

Key Practices for Designing Classroom Assessments with Social and Cultural Considerations
Dante Cisterna; Lei Liu, Educational Testing Service; Eowyn O'Dwyer, Educational Testing Service; David Baidoo-Anu
We will describe a set of assessment practices for classroom assessment grounded in sociocultural perspectives of learning. We identified three key practices for classroom assessment designers: (1) leverage student identity and funds of knowledge, (2) support student agency and empowerment, (3) promote civic awareness and social justice. We will provide examples of classroom assessments that highlight these principles and suggestions for assessment designers.

Discussants:
Stanley N Rabinowitz, **EdMetric LLC**

106. Challenges in Online Testing and/or Online Proctoring

Paper Session
9:50 to 11:20 am
Marriott: Floor 5th - Chicago Ballroom B/C

Chair:
Steve Ferrara, **HumRRO**

Participants:

Are the Item-level Effects of Online Proctoring Bi-Directional?
Paul Edward Jones, Pearson VUE; Liberty Munson, Microsoft; Xiaolin Wang, Pearson VUE
This study investigates whether modal DIF effects in an on-line versus traditional test center administration of a certification exam are truly bi-directional, or the result of unidimensional true DIF obscured by contrary artificial DIF under the Rasch model.

The Practical Investigation of LRP Test Form Performance at Late-pandemic Stage
Jihang Chen, Boston College; Yu-Lan Su, Ascend Learning
This study aims to evaluate test performance and item drifting for an assessment that transformed from its pre-pandemic CBT mode at test centers into an online Live-Remote-Proctoring testing mode during the pandemic. The form and item performance were evaluated, and the results indicate some noticeable item drifting pre- vs. post-pandemic.

The Journey of An Assessment Program from LOFT to LRP during the Global Pandemic
Yu-Lan Su, Ascend Learning
An assessment program expanded from LOFT at test centers to randomized linear forms through LRP mode and LOFT LRP within two years of pandemic. During the journey of evolution, the program faced various challenges, promptly launched a pilot study, and adapted to overcome technology and architecture difficulties to successfully transform.

Pursuit of group comparability: How will process data help?
Hongwen Guo, Educational Testing Service
Given the affordability and abundance of process data in educational measurement, researchers have been using them to provide richer understanding of score meaning and test-taking strategies to help teaching and learning. Yet, recent studies have shown that relationships among item responses and process data and the construct may vary across different student subgroups. Using process data, this study examined group comparability between two different testing modes (at test center versus at home) and between groups with or without technology-related disruption, to address the following research questions: Can process data help explain away group differences? Can we interpret test interruption the same way between the two test modes?

Detect the Impact of At-Home Testing Using a Pseudo-Equivalent Groups Approach
Jing Miao, Educational Testing Service; Yi Cao, Educational Testing Service; Michael E. Walker, Educational Testing Service
The proposed study uses real data to assess potential differential effects associated with test center (TC) vs. at-home testing via remote proctoring (RP). We use statistical approaches to balance the two subgroups of candidates choosing either testing option to detect the impact of at-home testing on participation and performance.

Discussant:
Susan Davis-Becker, **ACS Ventures, LLC**

107. Historical Perspectives on Educational Measurement

Coordinated Paper Session
9:50 to 11:20 am
Marriott: Floor 5th - Chicago Ballroom D

The purpose of this coordinated session is to acquaint educational measurement professionals and students with the historical underpinnings of current educational testing theory and practice. Its premise is that a clear understanding of the present and future of the field requires a solid understanding of its past and the trajectory by which we got to the present. The session consists of four papers focusing on 1) the history of equating, 2) the history of standard setting 3) the history of Bayesian inference in educational measurement, and 4) historical links between educational measurement, IQ testing, Eugenics, and the current anti-testing movement. It has been said that "We understand best those things we see grow from their very beginnings". These presentations help to show how educational measurement has grown from its beginnings and how its place in society and the perceptions of the public have developed over 120 years.

Session Organizer:
Brian Clauser, **National Board of Medical Examiners**

Participants:

History of Test Equating Methods and Practices Through 1985
Michael Kolen, The University of Iowa
A Brief History of Standard Setting
Mark Reckase, Psychometric Solutions
A History of Bayesian Inference in Educational Measurement
Roy Levy, Arizona State University; Robert J. Mislevy, Retired
The Lasting Impact of IQ-testing and the Eugenics Movement on Educational Measurement
Brian Clauser, National Board of Medical Examiners; Jerome Clauser, American Board of Internal Medicine; Amanda Clauser, National Board of Medical Examiners

Discussants:
Michael Kane, **Educational Testing Service**

108. Using New Techniques to Gather Validity Evidence

Paper Session
9:50 to 11:20 am
Marriott: Floor 5th - Chicago Ballroom E

Chair:
Guanlan Xu

Participants:

Utilizing NLP Techniques for Collecting Validity Evidence for a Situational Judgement Test
Okan Bulut, University of Alberta; Alexander MacIntosh, Altus Assessments; Cole Walsh, Altus Assessments

In this study, we aim to harness natural language processing (NLP) techniques for collecting construct validity evidence for a situational judgement test with constructed-response items. Using a large dataset from an operational SJT focusing on professionalism, we implement a semi-supervised NLP approach to explore the constructs underlying the test.

Using Eye Tracking to Validate Cognitive Processes in High Stakes Assessments
Jay Thomas, ACT, Inc.

This paper uses eyetracking data including heat maps and sequence maps to examine cognitive processes used on tasks from high-stakes assessments. Differences in gaze paths support claims that High, Mid, and Low scorers on a construct use different cognitive processes and possess different KSAs in assessments of math, reading, and science.

MxML: Exploring the Relationship Between Measurement and Machine Learning in Recent History
Yi Zheng, Arizona State University; Steven Nydick, Duolingo; Sijia Huang, Indiana University Bloomington; Susu Zhang, University of Illinois at Urbana-Champaign

The Zeitgeist of machine learning has transformed many disciplines, including educational measurement. Studies incorporating/discussing ML within measurement contexts have grown rapidly. As Phase I of the MxML project, this study systematically examines the latest 10 years' literature to explore the role ML has played in measurement research and applications.

Evaluating Content-related Validity of Mathematical Diagnostic Items Using a Topic Modeling Approach
Jiehan Li

This study proposes a new method to collect content-related validity evidence using the Latent Dirichlet Allocation to seek the topic distributions for test items. The study trains and compares two supervised machine learning models to classify the test items based on their topic distributions. Results include the prediction accuracy of two machine learning models compared to the classification by content experts using a publicly available dataset with 500 mathematical diagnostic items.

Discussant:
Richard Patz, **UC Berkeley**

109. Test Security

Paper Session
9:50 to 11:20 am
Marriott: Floor 5th - Chicago Ballroom F

Chair:
Katherine Nolan, **Curriculum Associates**

Participants:

Data Augmentation or Dimension Reduction in Machine Learning for Cheating Detection?
Dandan Liao, McKinsey & Company; Hong Jiao, University of Maryland

Data augmentation in machine learning is often recommended for situations where the data are scarce while dimension reduction is applied when the feature space is high-dimensional. This empirical study explores the application of data augmentation and dimension reduction in machine learning for cheating detection in large-scale assessments.

Using Item Scores and Distractors to Detect Item Compromise and Preknowledge
Kylie Gorney, University of Wisconsin-Madison; James Wollack, University of Wisconsin; Sandip Sinharay, Educational Testing Service; Carol Eckerly, Educational Testing Service

In this paper, we use item scores and distractors to detect compromised items and examinees with preknowledge simultaneously. Through simulations, we show that our method can detect preknowledge of a correct answer key and preknowledge of an incorrect answer key. A real data example is also included.

Aberrant Responding Detection in Multidimensional Forced-Choice Tests: lz vs. Optimal Appropriateness Measurement
Naidan Tu, University of South Florida; Lavanya Shravan Kumar, University of South Florida; Sean Joo, University of Kansas; Stephen Stark, University of South Florida

This research evaluated the efficacy of lz relative to optimal appropriateness measurement (OAM) methods for detecting aberrant responding on multidimensional forced choice tests based on the Multi-Unidimensional Pairwise Preference model. Simulation results indicated that lz performed as well or slightly better than OAM in all conditions.

Longitudinal Analysis of Response Accuracy and Time: Baseline Trends for Compromised Items
Merve Sarac, UW-Madison; Rich Feinberg, National Board of Medical Examiners; Chunyan Liu, National Board of Medical Examiners; Linette P. Ross, NBME

We investigated whether compromised items detected through external discovery methods differ in their response accuracy and time longitudinal trends compared to non-compromised items on a licensure examination. Using a purified sample near the cut score, we found no evidence that items detected through external discovery mechanisms provide a performance advantage.

Comparison of Likelihood Ratio Test-Based Preknowledge Detection Statistics for Use in Real-Time
Merve Sarac, UW-Madison; James Wollack, University of Wisconsin

We evaluated various likelihood ratio-based indices using different multiple comparison approaches for their potential to detect preknowledge in real-time forensics. Simulation results suggested that a maximum statistic based on the signed likelihood ratio test was the most promising combination studied.

Discussant:
Paulius Satkus, **Graduate Management Admission Council**

110. Using Measurement to Improve Educational Decisions

Paper Session
9:50 to 11:20 am
Marriott: Floor 5th - Chicago Ballroom G/H

Chair:
Anthony D. Fina, **University of Iowa**

Participants:

Using Value-Added Measures to Improve Academic Achievement Through Student-Teacher Matching
Peter Halpin, UNC-Chapel Hill; Matthew Springer, UNC; Christopher Brooks, UNC

The use of student test scores for teacher- and school-based accountability systems is now commonplace. This research seeks to move beyond accountability by using student test scores to match students to teachers in ways that optimize academic achievement.

It’s Complicated: Disproportionality when Using Multiple Measures to Select Students
Megan Welsh, University of California, Davis

Concerns have existed for some time with respect to the role of large-scale standardized tests in under-selecting minoritized students for honors (Atkinson, 2001). This study examines the extent to which multiple measures-based decisionmaking systems can be designed to amplify or mute the test bias detected in specific tests.

Investigating the Use of Measures for Mathematics Instructional Improvement
Marsha Ing, University of California, Riverside; Kara Jackson, University of Washington; Paul Cobb, Vanderbilt University; Thomas M. Smith, Vanderbilt University

We describe a validity approach to using measures designed to support district and school efforts to improve the quality of middle grades mathematics teaching. Findings suggest ways to leverage measurement for better decisions by systematically learning from their iterative use across different contexts.

Supporting Instructional Decisions with Group-level, Standard-Specific Inferences
Thomas Christie, NWEA; Carson Cook, NWEA; Garron Gianopulos, NWEA

To increase the actionability of information from existing adaptive standardized assessments, we propose a method for constructing class summaries of student performance at the granularity of individual standards. This method requires no change in assessment design, and we show via simulation the quality of inferences we produce about class parameters.

Risk of Bias for Validation Studies in Educational Measurement
Juyoung Jung, The University of Iowa; Ariel M. Aloe, University of Iowa

Educational measurement research is affected by potential sources of bias. Researchers collect validity evidence to assess quality within validation studies. Risk of bias should be formalized to identify unintended biases that yield systematic errors. Our systematic tool for assessing risk of bias would help to prevent drawing false inferences.

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

111. Advances in Item Response Modeling

Paper Session
9:50 to 11:20 am
Marriott: Floor 5th - Denver/Houston

Chair:
Yong He, **Measurement Incorporated**

Participants:

Robust Estimation of Latent Traits with the Graded Response Model
Audrey Filonczuk, University of Notre Dame; Ying Cheng, University of Notre Dame

A robust estimator to counteract aberrant responses in assessments containing Likert-type items is proposed. Simulations reveal the estimator reduces bias and MSE for different test lengths, numbers of response categories, and types of response disturbances, especially with the bisquare weighting system.

Moderating Adverse Impact Without Risking Selection Validity: Potential Application of IRTree Models
Victoria L. Quirk, University of Illinois at Urbana-Champaign; Justin L. Kern, University of Illinois at Urbana-Champaign

IRT models organized in a decision-tree structure (IRTree models) can model cognitive processes involved in response selection. Here, we investigate IRTree model selection validity and demonstrate the IRTree model's use in moderating adverse impact effects related to extreme response styles by reducing bias in measurement of a target trait.

Revisiting Benefits of Answer Change: The Role of Item Review
Weicong Lyu, University of Wisconsin - Madison; Daniel Bolt, University of Wisconsin, Madison

We propose an extension of an IRTree model by Jeon et al. (2017) to study the benefit of allowing answer change. By adding an item review stage and attending to item specific factors, we increase the plausibility of the model and demonstrate the sensitivity of the benefit function to assumptions regarding who reviews.

Performance of Effort-Moderated Multidimensional Item Response Model under Nonrandom Rapid-guessing Responses
Bowen Wang, University of Florida; Anne Corinne Huggins-Manley, University of Florida

Extending the effort-moderated model, an effort-moderated multidimensional item response theory model is proposed in this study. We found that nonrandom rapid-guessing patterns significantly affected the performance of models in estimating ability parameters. The multidimensional model outperformed the unidimensional model under conditions of a strong relationship between rapid-guessing and ability.

Extended Sequential Item Response Models for Multiple-Choice, Multiple-Attempt Test Items
Yikai Lu, University of Notre Dame; Ying Cheng, University of Notre Dame

An extension of the sequential IRT for multiple-choice multiple-attempt test items is proposed. Our new models have a freely-estimated pseudo-guessing parameter to accommodate different success rates of guessing, and have advantages over the previous models in having more possible shapes of response functions and being more likely to fit data.

Discussants:
Brian Habing, **National Institute of Statistical Sciences**

112. Standard Setting

Paper Session

9:50 to 11:20 am

Marriott: Floor 5th - Los Angeles/Miami

Chair:

Yi-Fang Wu, **Cambium Assessment, Inc.**

Participants:

ROC Validity Evidence of Angoff Cut Scores with Alternate Classification Criteria

Kari Hodge; R Noah Padgett, Baylor University

The Angoff method is widely used to establish defensible pass/fail scores for performance exams. Additional validity evidence for cut-scores can be gained with receiver-operating-characteristic (ROC) analysis. Using data from a credentialing program, results from ROC analyses indicated that cut-scores set by Angoff may be too low, implications are discussed.Using Eye Tracking to Validate Cognitive Processes in High Stakes Assessments

The Standard Setting Process for Equity Tools: Basics to Consider

Anica Bowe, Oakland University; Lynnette Mawhinney, Rutgers-Newark; Elizabeth Drame, University of Wisconsin-Milwaukee; Faith R Kares, Beloved Community; Carla Melaco, Beloved Community

This paper highlights issues raised by panelists during the Standard Setting Process for an Equity Audit tool. Our findings raise the questions on the extent to which DEI work could be measured, what the growth trajectories may look like, and the competency patterns for organizations engaging in DEI work.

Improving Developmental Appropriateness of Proficiency Level Descriptors for English Language Proficiency

Lynn Shafer Willner, University of Wisconsin - Madison

How might developmental appropriateness of proficiency level descriptors (PLDs) for English language proficiency standards and assessment be improved? This paper shares modeling, alignment (including linkages with CEFR), and validation activities to create six grade-level clusters of PLDs for the WIDA English Language Development Standards Framework, 2020 Edition.

Profile Selection, Variability, and Range in Standard Setting for Diagnostic Classification Models

Zachary Feldberg, University of Georgia

To use DCMs for federal accountability, their multidimensional, dichotomous results must be mapped onto ordinal scales. We experimentally examine the impact of the profile selection, variability, and attribute range used during a DCM standard setting by presenting sample student profiles to panels of experts and comparing the resulting cut points.

Discussant:

Michael R Peabody, **National Association of Boards of Pharmacy**

113. Predicting Item Difficulty and Response Latencies

Paper Session

11:40 to 1:10 pm

Marriott: Floor 5th - Chicago Ballroom B/C

Chair:

Priya Kannan, **WestEd**

Participants:

Predicting Item Difficulty from Students' Response Behaviors

Guher Gorgun, University of Alberta; Bin Tan, University of Alberta; Tarid Wongvorachan, University of Alberta; Okan Bulut, University of Alberta

The purpose of this study is to predict item difficulty in an adaptive learning environment by training two machine learning regressors (i.e., regression and neural network) with students' response behavior as features (e.g., response time, number of attempts). We achieved promising results with the model trained with a neural network regressor.

Predicting Response Latencies on Test Questions Using Qualities of the Written Text

Madelynn Denner, University of Notre Dame; Xiangyu Xu, University of Notre Dame; Teresa Ober, Educational Testing Service (ETS); Bo Pei, University of Notre Dame; Ying Cheng, University of Notre Dame

Can qualities of written test questions predict median response latencies? Data came from 610 items measuring AP Statistics domain knowledge. Analyses were conducted using machine-learning methods (stepwise forward and backward regression, Lasso regression, and a regression tree). Results have implications for estimating response latencies without student data.

Field-Testing Items Using Artificial Intelligence: Natural Language Processing with Transformers

Hotaka Maeda, Smarter Balanced

Two thousand variations of the RoBERTa model, an artificially intelligent "transformer" that can understand text language, completed an English literacy exam with 29 multiple-choice questions. Data were used to calculate the psychometric properties of the items, which showed some degree of agreement to those obtained from human examinee data.

The Impact of Cognitive Characteristics and Image-Based Semantic Embeddings on Item Difficulty
Michael Andreas Michels; Caroline Hornung, University of Luxemburg; Sylvie Gamo, University of Luxemburg; Pamela Isabel Inostroza Fernández, University of Luxemburg; Mark J Gierl, University of Alberta; Pedro Cardoso-Leite, University of Luxemburg; Antoine Fischbach, University of Luxemburg; Philipp Sonnleitner, Luxembourg Centre for Education

The impact of cognitive characteristics and semantic embeddings on item difficulty of 340 mathematics items is assessed across four grades (1, 3, 5, and 7). The study was conducted in the course of the Luxembourgish school monitoring program and has a total sample size of n = 19,799. Linear logistic linear logistic test models (Fischer 1978) are applied to investigate the effects. Cognitive characteristics can be mainly validated whereas changing semantic embedding mostly show effects in combination with cognitive factors.

Discussant:
Janet Mee, **NBME**

114. [SIGIMIE Session] Leveraging Process Data to Better Understand Engagement and Motivation in Large-Scale Assessment

Coordinated Paper Session
11:40 to 1:10 pm
Marriott: Floor 5th - Chicago Ballroom D

With the advent of computer-based testing, the types of data available to test developers and researchers has greatly expanded. So-called, process data, defined as “empirical data reflecting the course of working on an item” (Goldhammer & Zehner, 2017, p. 128) include, among other processes, the amount of time spent on an item and the number of actions (mouse clicks, keyboard entries, and the like). Beyond knowing whether an examinee selected (or provided) a correct answer and which of some fixed set of response options were chosen, we can also learn whether examinees lingered over an item or rapidly skipped it and how many actions (mouse clicks, page navigation) the examinee took until they moved on to the next item. This additional window into test-taking behavior is the focus of this coordinated session. Over four papers, we investigate several issues around test-taking motivation in a low-stakes setting with cutting-edge methods.

Session Organizer:
Leslie Rutkowski, **Indiana University**

Participants:

What Response Times Tell Us About Aalidity: Comparability and Engagement Issues in International Large-Scale Assessment
Maria Bolsinova, Utrecht University; Jesper Tijmstra, Tilburg University; Leslie Rutkowski, Indiana University; David Rutkowski, Indiana University

Test Engagement in Large-Scale Assessments Using Process Data and Unsupervised Learning Techniques
Hyo Jeong Shin, Educational Testing Service; Frederic Robin, ETS

Test Engagement and Multistage Adaptive Testing
Janine Buchholz, DIPF | Leibniz Institute for R; Hyo Jeong Shin, Educational Testing Service; Maria Bolsinova, Utrecht University

Evaluating Consistency of Behavioral Patterns Using Process Data in International Large-Scale Assessments
Qiwei He, Educational Testing Service

Discussants:
Francesco Avvisati, **OECD**

115. Putting Humpty Dumpty Back Together: Practical Advice for Synthesizing Validity Evidence

Organized Discussion
11:40 to 1:10 pm
Marriott: Floor 5th - Chicago Ballroom E

Cronbach (1971) famously wrote, “Construct validation is therefore never complete” (p. 452), which might have frightened users away from engaging in validation for fear of getting stuck in never-ending studies. Haertel (1999) emphasized that individual pieces of evidence do not make an assessment system valid or not. The validity evidence and logic must be prioritized and synthesized to evaluate the IUA. The process of integrating and synthesizing various types of empirical evidence and logical arguments into an evaluative judgment can be a challenging activity even if all the evidence was on the table. However, in operational assessment programs, new and varied information comes in over an extended time frame and is typically fragmented, especially the evidence related to consequences, thereby exacerbating the challenge. This session brings together the co-authors of the Validity and Validation chapter in the forthcoming Educational Measurement 5th Edition and authors of two recent articles on validity and validation to engage in a structured discussion focused on improving the quality and frequency of validity evaluation for our current state and interim assessments. Improved validation is a critical component in addressing the conference theme of “leveraging measurement for better decisions.”

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

Participants:
Scott Marion, National Center for the Improvement of Educational Assessment
Suzanne Lane, University of Pittsburgh
Michael Russell, Boston College
Daria Gerasimova, University of Kansas

116. Cognitive Diagnostic Assessment: Modeling and Design

Paper Session
11:40 to 1:10 pm
Marriott: Floor 5th - Chicago Ballroom F

Chair:
Magdalen Beiting-Parrish, CUNY Graduate Center

Participants:

A Construct Modeling Assessment Design Approach to Diagnostic Assessment
Joshua Sussman, UC Berkeley; Karen Draney, University of California, Berkeley; Perman Gochyyev, University of California, Berkeley

This paper focuses on the construct modeling (CM) approach to diagnostic assessment. First, we compare CM with other diagnostic assessment approaches. Then, we examine CM in depth using an operational, large-scale early childhood assessment developed using CM as an example. The discussion contains a critical synthesis and assessment design recommendations.

Pre-Assembly Methods for Cognitive Diagnostic Multistage Adaptive Testing
Jungwon Rachael Ahn; Leah Feuerstahler, Fordham University

Cognitive Diagnostic Multistage Adaptive Test (CD-MST) with module preassembly is a new test mode integrating the advantages of CDMs and MST. This study proposes holistic CD-MST preassembly methods to simultaneously consider statistical and non-statistical constraints by integrating an item selection method for CD-CAT and Liao and Jiao (under review)'s method.

Next-Generation CD-CAT: The Nonparametric Item Selection Method for Multiple-Choice Items
Yu Wang, University of Minnesota, Twin Cities; Chia-Yi Chiu, University of Minnesota

A novel nonparametric item selection method for multiple-choice items that combine the nonparametric classification method for multiple-choice items (MC-NPC) and the general nonparametric item selection (GNPS) method is proposed in the study. The preliminary study shows that the proposed method outperforms the GNPS method and reaches high agreement rates quickly.

Evaluating the Performance of Person-Fit Detection Methods in Diagnostic Classification Models
Jeffrey Hoover, University of Kansas; William Jacob Thompson, University of Kansas

We conducted a simulation to evaluate the performance of four machine learning models for identifying poor person-fit in diagnostic classification models. We compared the machine learning model performance with the performance of person-fit statistics and posterior predictive model checks. Performance was quantified as the Type I error and statistical power.

Validation of Diagnostic Classification Models for Diagnosing Misconceptions with Constructed-Response Items
Yuan Ge, The College Board; Louis Roussos, Cognia; Liuhan Sophie Cai, Cognia; La'Shea Cirlot, Cognia

This study aims to validate the generalized diagnostic classification models for multiple-choice option-based scoring (GDCM-MC; DiBello et al., 2015). We will explore the validity of the statistical classifications by using GDCM-MC with constructed-response items in the context of a performance assessment.

Discussant:
Hong Jiao, University of Maryland

117. Development and Methodologies for Operational CAT Programs with Advanced Requirements

Coordinated Paper Session
11:40 to 1:10 pm
Marriott: Floor 5th - Chicago Ballroom G/H

Computerized Adaptive Testing (CAT) has been increasingly used for statewide assessment in recent years. More than that, the requirements of a CAT test have also evolved rapidly to accommodate various needs of a state testing program, for example, the requirements of multiple administrations through the year, and the utilization of the items that require scoring from an artificial intelligence (AI) system. This symposium focuses on several of the challenging but important topics in the current trend of operational CAT programs. The first paper in this coordinated session presents Item Pool Development Recommendation (IPDR) software that can provide directions for item writers to target on writing the most needed items. The second paper introduces a flexible item calibration algorithm that better suits the unique requirements of adaptive formative assessment. The third paper compares some adaptive-within-passage algorithms for passage-based CAT tests. Neither method compared requires pre-assembly of testlets and thus will save much time in the preparation of tests. The fourth paper investigated how AI-scored items should be inserted in CAT with respect to the location and the number of AI items inserted and its impact to the CAT performance.

Session Organizer:
Yang Lu, **Pearson**

Participants:
Design an MST Item Pool Development Recommendation Software
Yuehmei Chien
Item Calibration Design for Adaptive Formative Assessment
Hao Ren, PEARSON
A Comparison of Within-Passage Item Selection Methods
Yu (Tracy) Zhao; Steve Fitzpatrick, Pearson
Computerized Adaptive Testing with AI-Scored Items
Haiyan Lin, Pearson; Yang Lu, Pearson; David Shin, Pearson

Discussant:
Shiyu Wang

118. Causal Modeling of Log Data from EdTech

Coordinated Paper Session
11:40 to 1:10 pm
Marriott: Floor 5th - Denver/Houston

Most educational technology (EdTech) applications automatically document every action every learner takes while using the software. Our goal in this session is to use this log data to discover how students learn best. The four papers combine psychometrics, statistics, and machine learning and discuss four of the roles log data can take in causal research–as covariates, outcomes, the intervention itself, or as mediators or moderators. The first two papers use data from randomized trials on EdTech platforms. The first discusses the use of log data from both students who participated in the experiment as well those who didn't to estimate more precise impacts. The second describes the use of log data to define experimental outcomes that provide more nuanced measurements of learning. The third paper describes a regression discontinuity to study the effect of a software's reward structure on students' implementation. The fourth paper uses log data from a randomized field trial of an EdTech product to summarize patterns in which students or teachers implement the program, and then estimates how treatment effects vary between different usage patterns. The session will conclude with a discussion led by Professor Neil Heffernan, a leader in the development, study, and experimentation within EdTech.

Session Organizer:
Adam C Sales, **Worcester Polytechnic Institute**

Participants:
EdTech A/B Testing using Auxiliary Log Data and Deep Learning
Adam C Sales, Worcester Polytechnic Institute; Ethan B Prihar, Worcester Polytechnic Institute; Johann Gagnon-Bartch, University of Michigan; Neil T Heffernan, Worcester Polytechnic Institute

We present unbiased causal estimates coupling design-based causal estimation to machine-learning models of log data from users who were not in the experiment from over 250 randomized A/B comparisons. Incorporating auxiliary data into causal estimates can be equivalent to increasing the sample size by as much as 50-80%.

Identifying a Surrogate Measure of Long-Term Learning
Ethan B Prihar, Worcester Polytechnic Institute

Learning platforms conducting A/B tests often measure learning with metrics such as whether students correctly answered the next problem on their first try with no further support. We used data from assignments with post-tests to identify a measure of learning based on log data that correlates with post-test scores.

Evaluating In-Program Decisions by Leveraging Regression Discontinuity Analysis for Causal Inference
Kirk P Vanacore, Worcester Polytechnic Institute; Erin Ottmar, Worcester Polytechnic Institute; Adam C Sales, Worcester Polytechnic Institute; Allison Liu, Worcester Polytechnic Institute

We evaluate a feedback system in the From Here to There! online math tutor with regression discontinuity. Students receive one to three clovers depending on the number of steps they took to complete a problem. We find that receiving one clover, rather than two, causes students to replay the problem.

Bayesian and Maximum Likelihood Estimation in Fully Latent Principal Stratification
Tiffany Whittaker, University of Texas; Hyeon-Ah Kang, University of Texas at Austin; Sooyong Lee, University of Texas; Adam C Sales, Worcester Polytechnic Institute

In a randomized field trial of EdTech, principal stratification can find different treatment effects for different modes of implementation. Fully latent principal stratification (FLPS) uses measurement models to extend this method to complex multivariate implementation (log) data. We contrast Bayesian and maximum likelihood approaches to FLPS estimation with simulation studies.

Discussant:
Neil T Heffernan, **Worcester Polytechnic Institute**

119. Advancing Psychometric Processes and Tools in a Changing Testing Environment

Coordinated Paper Session
11:40 to 1:10 pm
Marriott: Floor 5th - Los Angeles/Miami

With the changing testing environment (such as remote testing, adaptive testing) and increasing emphasis on equity and fairness, existing practice needs to be examined and improved, new methodologies need to be developed and implemented, and operational processes need to be streamlined. The set of five papers in the session addresses some of the measurement challenges we are facing in practice: tackling small sample issues using advanced psychometric models in adaptive testing; reexamining item analysis practice by incorporating distractor analysis quantitatively and qualitatively; simplifying complex models for practical implementation to detect preknowledge; and developing new item fit statistics that incorporating item parameter estimation error.

Session Organizer:
Hongwen Guo, **Educational Testing Service**

Participants:
Did A Distractor Function Well in a Multiple-Choice Item?
Yi Cao, Educational Testing Service; Hongwen Guo, Educational Testing Service; Youhua Wei, Educational Testing Service; Kathryn Hille, ETS;
Gautam Puhan, ETS
A Hierarchical Bayesian Approach to Small-sample Item Calibration: Added-values and Additional Steps
Ikkyu Choi; Yi Cao, Educational Testing Service; Zhuangzhuang Han, ETS; Hongwen Guo, Educational Testing Service
Connecting Item Statistics and Content to Improve Test Development and Learning
Kathryn Hille, ETS; Youhua Wei, Educational Testing Service; Hongwen Guo, Educational Testing Service; Yi Cao, Educational Testing Service;
Gautam Puhan, ETS
Towards Practice: A Simplified Variant of the Two-way Outlier Detection Model for Pre-knowledge Detection
Zhuangzhuang han, ETS; Sandip Sinharay, Educational Testing Service
Fit Statistics for Item Response Functions Based on Generalized Residuals
Xiangyi Liao; Peter van Rijn, ETS Global; Sandip Sinharay, Educational Testing Service

Discussant:
Jonathan Weeks, **Educational Testing Service**

120. Comparability of Scores from Through-Year and Traditional State Assessments: Examining Louisiana

Coordinated Paper Session
1:30 to 2:30 pm
Marriott: Floor 5th - Chicago Ballroom B/C

In 2018 and 2019, under the Innovative Assessment Demonstration Authority (IADA), the Office of Elementary and Secondary Education invited states to submit applications to establish and operate innovative assessment programs as alternatives to existing end-of-year state summative assessments. Four states received approvals: Georgia, North Carolina, New Hampshire, and Louisiana. In Louisiana, students historically have taken the Louisiana Educational Assessment Program (LEAP) test. Under IADA, however, Louisiana began developing and piloting an innovative through-year, curriculum-embedded English Language Arts assessments, known as the Innovative Assessment Program (IAP). Louisiana will be one of the first states to report summative scores of record from an innovative state assessment, beginning with grade seven students that participated in the IAP pilot in school year 2021-22. With grade seven students across Louisiana receiving scores from two different ELA assessments (i.e., LEAP and IAP), it is critical to understand how scores from each assessment are comparable. This session presents three papers that address the overarching research question: what evidence exists to demonstrate the degree of comparability of IAP and LEAP scale scores? The papers focus on the scaling methodology, empirical results, and educator ratings of alignment between IAP content and LEAP achievement levels.

Session Organizer:
Audra Kosh, **NWEA**

Participants:

Methodological Decisions in Scaling Louisiana's Through-Year Assessment
Nathan Dadey, Center for Assessment; David Hopkins, Louisiana Department of Education; Xiangdong Liu, University of Iowa; Leslie Keng, Center for Assessment; Audra Kosh, NWEA; Shudong Wang, NWEA

Empirical Results for Comparability of Louisiana's Through-Year Assessment
Audra Kosh, NWEA; Nathan Dadey, Center for Assessment; David Hopkins, Louisiana Department of Education; Xiangdong Liu, University of Iowa;
Leslie Keng, Center for Assessment; Shudong Wang, NWEA

Educator Judgements of Alignment of Louisiana's Through-Year Assessment
Nathan Dadey, Center for Assessment; David Hopkins, Louisiana Department of Education; Ruth Calliouet, Louisiana Department of Education;
Leslie Mugan, NWEA

Discussant:
Stephen G Sireci, **University of Massachusetts, Amherst**

121. Impact of College Admission Test Mandate and Alternative Approaches

Paper Session
1:30 to 2:30 pm
Marriott: Floor 5th - Chicago Ballroom D

Chair:
Stella Kim, **University of North Carolina at Charlotte**

Participants:

Concerns of Grade Inflation in High School Subject Grades Since 2010
Edgar Sanchez, ACT

Grades are meant to indicate students' academic knowledge, skills, and college and career preparation. Grade inflation raises concerns about the meaning and interpretation of grades. The present study makes use of a variety of methodologies to examine grade inflation in English, mathematics, social studies, and science from 2010 through 2021.

Validity of GPA Mathematics for Decisions on University Admission in Chile
Xaviera Gonzalez-Wegener, UCL Institute of Education

In Chile, GPA is used to inform the algorithm for university admission. By articulating Critical Disability Studies and Critical Realism frameworks, this qualitative study shows that a multilevel achievement disabling system that significantly affects the inferential validity of the GPA mathematics scores.

Discussant:
Wayne J. Camara, **LSAC**

122. 2023 NCME Career Award Session

Paper Session
1:30 to 2:30 pm
Marriott: Floor 5th - Chicago Ballroom E

Chair:
Robert Henson, **University of North Carolina - Greensboro**

Participants:

From Martingales to Formative Assessments (FAs): A Career in Progress
William Stout, Professor (Emeritus), University of Illinois at Urbana-Champaign (Department of Statistics). And:
Research Professor (Emeritus) University of Illinois at Chicago (Learning Sciences Research Institute)

After obtaining my 1967 Mathematics Ph. D., my entire career was at the University of Illinois. After my epiphany that I wanted my research to be "applicable", in the '90s I switched from doing "pure" probability research (concerning the law of the iterated logarithm for weakly dependent random variables: aesthetically beautiful but so very unapplicable) to developing theoretically-grounded and, hopefully, useful educational measurement methodologies. Three foci resulted: assessing latent dimensionality, detecting test bias, (both described briefly) and, currently, using IRT-grounded diagnostic classification modeling (DCM) to improve classroom FAs (defined as assessments improving near-future teaching and learning).

One result of my switch, I formed the Statistical Laboratory for Educational and Psychological Measurement, 16 Ph. Ds. from three academic departments resulting. Their combined contributions to our field dwarf my contributions. For example, four succeeded me as presidents of the Psychometric Society. Currently, 11 hold university faculty positions here and abroad.

The bulk of my talk addresses DCM-grounded FA accomplishments by colleagues and me. Our Generalized DCMs (GDCMs) provide well-fitting models of MC/(CR-constructed response) item-based tests scored respectively at the MC-option/(CR-popular answer) level to diagnose examinee skills and misconceptions (two kinds of latent attributes). Our resulting methodologies should improve classroom and online student learning through much-improved diagnosis of examinee attributes.

123. Demonstrations: Session 1

Demonstration Session

1:30 to 2:30 pm

Marriott: Floor 5th - Chicago Ballroom F

Chair:

Joseph Grochowalski, **College Board**

Participants:

ITEMS: The Benefits of Two Major Changes for Learners and Authors

Brian C Leventhal, James Madison University

This demonstration will present two major changes to ITEMS and outline the benefits for learners and authors. These include benefits to educators who use ITEMS in their courses, professionals who want to refresh or learn new methods, and stakeholders of assessment. Additional focus will be on the simplified authorship process.

An AI Approach to Suggest Sampling Weights for ECLS-K Data Analysis

Huade Huo, AIR; Paul Bailey, American Institutes for Research; Ting Zhang, American Institutes for Research; Emmanuel Sikali

Sampling weight selection is an important step when analyzing longitudinal data but potentially overwhelming due to substantial number of available weights and complex rules. Using the ECLS-K:2011 as a pilot, the suggestWeights, an artificial intelligence function built in EdSurvey R package, assist analysts selecting sampling weights in their data analysis.

Empowering Untapped Talent with Comprehensive Assessment and Training for Workforce Success

Ou Lydia Liu, ETS; Kevin Williams, Educational Testing Service; Guangming Ling, Educational Testing Service

The U.S. workforce has a talent shortage and access to nontraditional talent is urgently required. Apprize is a platform that empowers untapped talent through job interest, skills, and behavioral competency assessment. Apprize connects talent with employers or with mission-driven training providers that can help improve their technical and non-technical skills.

Reducing Error in Test-Taker Classification with ATA

Jon Lehrfeld, Educational Testing Service; Yong Luo, NWEA

We demonstrate a method using automated test assembly (ATA) to reduce test-taker misclassification during test form assembly. This source of error, which stems from necessary and practical steps taken when producing conversion tables at the end of a test administration, can be mitigated using a relatively simple ATA procedure.

124. Innovations in Assessment and Feedback

Paper Session

1:30 to 2:30 pm

Marriott: Floor 5th - Chicago Ballroom G/H

Chair:

Yuan-Ling Liaw

Participants:

Automated Diagnosis and Feedback on Three-dimensional Science Reasoning

Lei Liu, Educational Testing Service; Dante Cisterna-Alburquerque, ETS; Yi Qi, Educational Testing Service; Devon Kinsey, Educational Testing Service; Kenneth Steimel, ETS

This presentation describes a proof-of-concept study that developed an assessment innovation with automated identification and feedback on student reasoning patterns related to multidimensional science learning. It also reports a cognitive interview study that investigates how the innovation impacted student performance and how students reacted to the automated diagnosis and feedback.

Conversation-based Assessments: Enhancing Students' Experiences of Formative Assessment via Human-like Interaction

Seyma N. Yildirim-Erbasli, Concordia University of Edmonton; Okan Bulut, University of Alberta; Ying Cui, University of Alberta

In this study, we designed a conversation-based assessment (CBA) for an undergraduate-level course with the goal of improving the formative assessment environment. We found highly accurate dialogue moves within CBA and positive student attitudes toward CBA. Our study suggests the utility of CBA in measuring knowledge as well as enhancing assessment experience.

Game-based Creative Problem Solving Assessment with Automatic Scoring and Multiple-board Equating

Xinchu Zhao, Roblox Corporation; Matthey Emery, Roblox Corp; Erica Snow, Roblox Corp; Jack Buckley, Roblox Corp

This study introduces a novel game-based creative problem-solving assessment built within Roblox that automatically scores patterns of behaviors extracted from telemetry data. We applied a simple-structure multidimensional item response theory (SS-MIRT) model in assessment scoring and equating to measure two different constructs: ideation and divergent thinking in CPS.

Discussant:

Kristen Huff, **Curriculum Associates**

125. The Development and Utility of Learning Progressions in the K-12 Setting

Coordinated Paper Session
1:30 to 2:30 pm
Marriott: Floor 5th - Denver/Houston

Academic standards across K-12 include rigorous expectations, represented as complex cognitive engagement with disciplinary ideas. Learning progressions emphasize a concomitant development of increased sophistication of thinking across the grades. For a learning progression to be useful in the classroom, it should meet the requirements defined by academic standards, use instructionally relevant skills for educator decision-making, and reflect empirical data used to validate it. However, Interpretation of scores on educational assessments typically relies on item difficulty, which is more widely studied and easier to operationalize empirically than are conceptualizations of rigor, complexity, or sophistication. Sorting out how these entangled ideas (of rigor, complexity, sophistication, difficulty, etc.) converge and diverge may help the field attain shared goals related to use of assessment scores to provide instructional guidance as relates to learning progressions. This session describes the instructional utility of a learning progression, how to develop such a learning progression, and a process of conducting a psychometric validation of the learning progression linked to an assessment that is used for screening and progress monitoring student learning.

Session Organizer:
Catherine Close, **Renaissance Learning**

Participants:
Empirical Validation of Learning Progressions and Linking Test scores to Inform Instruction
Catherine Close, Renaissance Learning
Development of Learning Progressions in Context of Academic Standards
Julianne Robar, Renaissance Learning
Differentiating Difficulty from Complexity to Promote Intended Uses of Learning Progressions
Sara Christopherson, Wisconsin Center for Education Products & Services- UW Madison

Discussant:
Amelia Gotwals, **Michigan State University**

126. Establishing Instructionally Meaningful Cut Scores with Embedded Standard Setting

Coordinated Paper Session
1:30 to 2:30 pm
Marriott: Floor 5th - Los Angeles/Miami

Instruction is best supported when assessment information is validly provided at the subscore level. Overall scores are useful but the instructional value for teaching is better when information is provided not just for the total score but also for subscores (e.g., Algebra, Geometry). When score interpretation needs to occur at a more than one level (e.g. total and subscore), a more granular approach to standard setting may be needed. This session will review how Embedded Standard Setting (ESS; Lewis & Cook 2020) might be used to establish performance levels for subscores which are then aggregated to the total score level using the i-Ready Diagnostic as an illustration. The primary objective of the i-Ready Diagnostic is to provide teachers with instructionally actionable information based on where each student is along the learning trajectory. To accomplish this the i-Ready Diagnostic generates scale scores and associated performance levels at both the overall subject level and for a set of content area domains. The Presenters show how subscore information can be supported by carefully crafting performance level descriptors and establishing cut scores at the subscore level via ESS. We will also present the results of a cut score validity study using instructional outcomes.

Session Organizer:
Laurie Davis, **Curriculum Associates**

Chair:
Laurie Davis, **Curriculum Associates**

Participants:
Aligning Assessment and Instruction through PLD Creation
Amanda Brice, Curriculum Associates
A Bottom Up Approach to Instruction and Assessment: Supporting Validity with ESS
Daniel Lewis, Creative Measurement Solutions LLC
Using Instructional Outcomes to Validate Cut Scores
Ted Daisher, Curriculum Associates; Kristin M. Morrison, Curriculum Associates

Discussant:
Jade Caines Lee, **University of Kansas**

127. Reception for Researchers from Historically Marginalized Groups

Plenary Session
1:30 to 2:30 pm
Marriott: Floor 7th - Salon I

The Historically Marginalized Groups (HMGs) mixer is a semi-structured social networking session for both graduate students and professionals that identify themselves with one or more groups that are underrepresented or historically marginalized in the field, and their allies/co-conspirators. Organized by the Diversity Issues & Testing Committee, the Graduate Student Issues Committee, and the Membership Committee, the event will include a variety of socially engaging group activities as well as light refreshments and beverages.

128. Challenges in Growth Measures and Accountability Decisions

Paper Session
2:50 to 4:20 pm
Marriott: Floor 5th - Chicago Ballroom B/C

Chair:
Xia Mao, **NBOME**

Participants:

Data-Rich/Information Poor: A Critique of NWEA MAP's Achievement and Growth Reportage
Paul Zavitkovsky, Center for Urban Education Leadership, University of Illinois at Chicago
Many school districts now invest heavily in proprietary assessment systems to monitor academic progress. MAP Growth is the most widely used of these systems. But data comparisons from Chicago illustrate troubling inconsistencies between MAP data reports and comparable reports derived from NAEP and other trusted sources.

Mitigating Students and Schools Evaluation Biases Due to Variation in Instructional Exposure
Yeow Meng Thum, NWEA
Although it is unfair to compare the performance of students with different levels of OTL, confounding effects due to student differences in instructional exposure stemming from known variation in district calendars and student testing schedules are routinely ignored. Appropriate comparisons employing projected performance from prior predictive distributions are examined.

A Classification of State CSI Exit Criteria
Michael Fienberg, University of Southern California Rossier School of Education
In response to ESSA requirements, states have designed varying criteria for states to exit Comprehensive Support and Improvement status. This study reviews all 51 states' exit criteria and classifies states into five categories and one subcategory, ranking them on the perceived difficulty for schools to exit CSI status.

Use of Assessment Results in Presence of Model Misspecification and Measurement Error
Salih Binici; Yachen Luo
This study examines consequences of model misfit and measurement error on reporting outcomes for a large-scale assessment. It investigates whether ignoring model misspecification and measurement error has any practical impact on reported scale scores for parents and teachers, also their secondary use in statistical analyses to inform policy makers.

An Investigation into Performance Mobility Using a Simulated State Accountability System
Michael Fienberg, University of Southern California Rossier School of Education
What would be the empirical and design features of an ESSA compliant system that would maximize the chances that school improvement would result in exiting low-performing designations? This study uses simulations to answer this question and recommends six principles for states to consider when designing their accountability systems.

Discussants:
Joshua Sussman, **UC Berkeley**

129. Rater Effect Evaluation and Mitigation

Paper Session
2:50 to 4:20 pm
Marriott: Floor 5th - Chicago Ballroom F

Chair:
Brad Thiessen

Participants:

Many-Facet Rasch Designs: How Should Raters be Assigned to Examinees?
Christine DeMars, James Madison University; Yelisey A. Shapovalov, James Madison University; John D Hathcoat, James Madison University
In Facets models, raters should be connected. Keeping the number of ratings constant, we found that the standard error of both rater severity and examinee ability was higher when raters scored one examinee in common with many different raters than when they scored many examinees in common with two raters.

Using Ordinal Rescore Measures to Monitor Rater Drift
John Donoghue, Educational Testing Service; Adrienne Sgammato, ETS

When re-using constructed response items, a set of Time A papers, rescored at Time B, are used to determine whether scorers are functioning differently. A new statistic is proposed that accounts for the ordinal nature of the scores and the rescore data's sampling structure. Simulation results support the measure's accuracy.

Using Iterative Generalizability Studies in the Context of Measuring Equitable Mathematics Instruction
Elizabeth L. Adams, Southern Methodist University; Anne G. Wilhelm, Southern Methodist University; Rachael N Becker, Southern Methodist University; Jonee Wilson, NC State University; Templ A. Walkowiak, NC State University

This iterative generalizability and decision study examines the stability of classroom observational scores across raters before and after revising rubrics. We examine data from 118 teachers rated by different subsets of 5 raters. Using two lessons and three raters yielded optimal results and stability increased for most rubrics following revision.

Leveraging Within-year Data for Trend Rescore using New Monitoring Statistics
John Donoghue, Educational Testing Service; Adrienne Sgammato, ETS

Scores produced by human trend scoring (rescoring Time A responses at Time B) of constructed response items are compared using procedures that assume a product-multinomial model, as opposed to the typical multinomial distribution. This simulation study compares distributional properties of three conditional statistics, with paired-t and Stuart's Q.

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

130. Expanding the Conceptualization of Fairness for Digital Learning and Assessment

Coordinated Paper Session
2:50 to 4:20 pm
Marriott: Floor 5th - Chicago Ballroom G/H

Learning and assessment have largely shifted to digital platforms which changes the playing field for fairness. Fairness in learning and assessment has been widely addressed in validity standards (AERA et al., 2014) and frameworks (Xi, 2010; Chapelle et al., 2008; Kunnan, 2000; Kane, 1992). Some fairness considerations addressed in these standards and frameworks (such as construct definition, and fairness and bias review) remain relevant for digital learning and assessment (DLA) systems. However, we need to expand our conceptualization of fairness. This includes, for instance, greater attention to diversity and inclusion for building AI for a more diverse set of learners and test takers, and alignment between education problems and machine learning solutions. Further, we need to develop mechanisms similar to algorithmic auditing of AI systems (Raji et al, 2020) in order to systematically audit the expansion of fairness issues in DLA. Through five presentations, this session discusses a DLA ecosystem for auditing, and exemplifies fairness considerations across design, measurement, security, and learner and test-taker experience and sociocognitive factors as we build DLA systems. Presentations have implications for developing modern methods (Johnson et al, 2022), frameworks (Huggins-Manley et al, 2022), standards, and policy that impact responsible AI in education (Dignum, 2021).

Session Organizers:
Jill Burstein, **Duolingo**
Geoff LaFlair, **Duolingo**
Alina A. von Davier, **Duolingo**

Participants:

Fairness Auditing in a Digital-First Learning and Assessment Ecosystem
Jill Burstein, Duolingo; Geoff LaFlair, Duolingo; Alina A. von Davier, Duolingo

Embracing Diversity and Inclusion While Accelerating Learning and Monitoring Growth at BrainPOP
Yigal Rosen, BrainPop; Barbara Hubert, BrainPop; Sara Bakken, BrainPop; Melissa Hogan, BrainPop

Designing Image-based Items for Cross-Cultural Assessments
Lisa Keller, University of Massachusetts

Equity in Learner and Test-Taker Experience
Maria-Elena Oliveri, University of Nebraska; Anson Green, Tyson Foods

Reimagining the Life Cycle of Machine Learning in Education
Lydia T. Liu, Cornell University; Serena Wang, UC, Berkeley; Tolani Britton, UC, Berkeley; Rediet Abebe, UC, Berkeley

Discussants:
Stephen Sireci, **University of Massachusetts Amherst**

131. Advances in Language Assessment

Paper Session
2:50 to 4:20 pm
Marriott: Floor 5th - Denver/Houston

Chair:
Jennifer Kobrin, **ATLAS: University of Kansas**

Participants:

Predicting Oral Reading Fluency Scores by Silence Time with Natural Language Processing
Yusuf Kara, Southern Methodist University; Akihito Kamata, Southern Methodist University; Emrah Emre Ozkeskin, Heilbronn University, Germany; Xin Qiao, Southern Methodist University

This study investigates evidence for meaningful pauses that are believed to be made by fluent readers in the context of oral reading fluency (ORF) assessments. We derive various text features from passages read by students and use relative silence times related to these features as predictors of ORF scores.

Developing Items for a Silent Reading Efficiency Task
Amy Burkhardt, Stanford University; Maya Yablonski, Stanford University; Jamie Mitchell, Stanford University; Liesbeth Gijbels, University of Washington; Jason Yeatman, Stanford University

A silent sentence reading efficiency test for progress monitoring requires a large item bank. We propose guidelines for developing items, conduct a case study to explore the utility of sentence-level statistics for refining an item bank, and explore the use of models to predict these statistics.

Estimating Item Difficulty: What Variables Do Subject Matter Experts Use?
Ayfer Sayin; Okan Bulut, University of Alberta

This study aims to determine the variables used by SMEs in estimating item difficulty in a Turkish test consisting of reading comprehension, reasoning, and grammar items. The goals of the research are: (i) to determine the variables that explain item difficulty prediction used by SMEs and (ii) to provide evidence to create a source for an automatic estimation model.

ACTFL Chinese Reading Proficiency Guidelines: Verifying the Difficulty Hierarchy
Jia Lin, Howard University

The American Council on the Teaching of Foreign Languages (ACTFL) Proficiency Guidelines have been guiding foreign language teaching, testing, textbook development. However, the guidelines have been criticized because the descriptors were not scaled empirically. This study aimed to empirically verify the difficulty hierarchy posited by the ACTFL Chinese Proficiency Guidelines-Reading.

Discussants:
Ye Tong, **National Board of Medical Examiners**

132. Data-Driven Analysis of Latent Structures for Cognitive Diagnosis Models in Educational Assessments

Coordinated Paper Session
2:50 to 4:20 pm
Marriott: Floor 5th - Los Angeles/Miami

Cognitive diagnosis models are a family of restricted latent class models that have been widely applied in educational assessments. The applications of these models rely on the specifications of two important latent structures. The first is the item-attribute Q matrix, which indicates the sets of attributes required by each item. The second is the attribute hierarchy, which specifies the relationships among the underlying attributes. These two components are typically provided by content exports or test developers. In this session, we will present several innovative data-driven approaches for statistical analysis on the latent structures of cognitive diagnosis models from response data. The Presenters will investigate the statistical inferences of attribute hierarchy under various model specifications or in the presence of testlets items. Data-driven approaches of learning Q matrices from hierarchical models or based on nominal responses will also be discussed. Results from these studies are expected to help educators and applied researchers to incorporate the information discovered from the data set to seek new insights into the cognitive theory and promote new developments for educational assessments and instructional interventions.

Session Organizers:
Shiyu Wang
Yinghan Chen, **University of Nevada, Reno**

Chair:
Shiyu Wang

Participants:

Bayesian Inference of Attribute Hierarchy in Cognitive Diagnosis Models
Yinghan Chen, University of Nevada, Reno; Shiyu Wang
Learning Latent and Hierarchical Structures in Cognitive Diagnosis Models
Chenchen Ma, University of Michigan; Jing Ouyang, University of Michigan; Gongjun Xu, University of Michigan
A Testlet Diagnostic Classification Model with Attribute Hierarchies
Wenchao Ma, University of Alabama; Chun Wang, University of Washington; Jiaying Xiao, University of Washington
Restricted Latent Class Models for Nominal Response Data: Identifiability and Estimation
Ying Liu, University of Illinois at Urbana-Champaign; Steven Culpepper, University of Illinois at Urbana-Champaign

Discussant:
Chun Wang, **University of Washington**

133. Business Meeting and Presidential Address

Business Meeting
4:40 to 6:15 pm
Marriott: Floor 5th - Chicago Ballroom D/E

134. President's Reception

Plenary Session
6:30 to 8:30 pm
Marriott: Floor 7th - Salon I and II

135. [Joint Session with AERA Division D] Examining AI and Machine Learning Through a Fairness and Equity Lens

Organized Discussion

8:00 to 9:30 am

[AERA Hotel] InterContinental Chicago Magnificent Mile: Floor 4th - Camelot Room

In this jointly organized AERA Division D and NCME session, assessment experts in the use of AI/ML applications and tools will have a structured panel discussion on how to address fairness and equity issues when using these tools. The conversation will examine process data, scenario- and game-based assessments, innovations in training scoring engines, and relevant AI learnings from adjacent fields such as computer science, engineering, and data science. Panelists will delve deeply into associated fairness concerns and mitigating strategies and on the impact of AI on developing cogent validity arguments and in making better decisions based on examinee scores.

Moderator:
Mary Pitoniak, **ETS**

Presenters:
David Dorsey, **HumRRO**
Jack Buckley, **Roblox Corp**
Kadriye Ercikan, **Educational Testing Service**
Steve Ferrara, **HumRRO**

136. Combining Innovation and PAD to Economize Assessment Processes that Support Better Decisions

Coordinated Paper Session

8:00 to 9:30 am

Marriott: Floor 5th - Chicago Ballroom B/C

Principled Assessment Design (PAD) is intended to guide assessment development efforts such that assessments elicit evidence specifically relevant to claims about what the assessment is measuring. Range Achievement Level Descriptors (RALDs; Egan et al., 2012), which are an important component of many PAD approaches, are intended to reflect the evidence-based range of content- and cognitive-skills associated with student growth and sophistication on increasingly difficult construct-relevant tasks. They are intended to be cognition-based within-grade learning progressions that are critical for communicating how tasks align to standards and how they elicit evidence of a student's stage of learning. Intentionally and consistently embedding item-to-RALD alignment throughout the assessment development lifecycle provides a unifying framework that enhances the instructional utility of the information provided to teachers to support them in making better decisions. Because there is some concern that PAD can make assessment development more costly, this session focuses on implementation of PAD processes for item-to-RALD alignment that are innovative and can also be automated for cost savings across the item development to score reporting phases. We will discuss and illustrate progressive yet cost saving processes leading to a reporting framework that provides personalized, efficacious information about what students know and can do.

Session Organizer:
Christina Schneider, **Cambium Assessment, Inc.**

Moderator:
Kevin Dwyer, **Cambium Assessment**

Participants:
Automating Item Specifications from Range ALDs
Christina Schneider, Cambium Assessment, Inc.; Jing Chen, Cambium Assessment; Margaret McMahon, Cambium Assessment
Automating the Estimation of Cut Scores via Embedded Standard Setting
Daniel Lewis, Creative Measurement Solutions LLC; Robert Cook, Cognia
Aligning Rubrics and Scoring to RALD-based-Assertions to Automate Feedback to Students
Susan Lottridge, Cambium Assessment; Kevin Dwyer, Cambium Assessment; Ben Godek, Cambium Assessment
Using RALDs to Generate Actionable Reporting Statements
Ellen Forte, edCount, LLC; Melissa Fincher, edCount, LLC

Discussant:
Richard Melvin Luecht, **University of North Carolina at Greensboro**

137. Automated Test Assembly in Operational Assessment Programs

Coordinated Paper Session
8:00 to 9:30 am
Marriott: Floor 5th - Chicago Ballroom D

The purpose of this symposium is to review the research and technical challenges that have been faced with building and maintaining the automated test assembly (ATA) tools in different operational testing programs. The first presentation (Diao) will focus on model building and form assembly for an assessment program that measures Interpersonal and intrapersonal skills for high-stakes higher education admission. The second presentation (Lehrfeld) will focus on a state-wide testing program, the challenges the psychometrician and assessment development team faced during the first year of ATA setup and implementation, and how they provided sound solutions to those challenges. The third presentation (Yoo & Fu) will discuss the ATA challenges and solutions for a large-scale assessment with an adaptive testing design. This presentation will discuss the importance of balancing the three types of ATA specifications (e.g., content, statistical, and test security) to build multiple parallel test forms with high-quality. The fourth presentation (Li, Li, Manna & Gu) will focus on assembly challenges for a newly-developed large-scale English assessment. And the fifth presentation (Lim & Han) will introduce a new pool assembly framework that helps build many parallel item pools of Graduate Management Admission Test (GMAT) efficiently.

Session Organizers:
Qi Diao, **ETS**

Moderator:
Qi Diao, **ETS**

Participants:

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

The presentation will focus on model building and form assembly for an assessment program that measures interpersonal/intrapersonal skills for high-stakes higher education admission. This presentation will discuss how to mixed integer programming to create statement blocks (pairs and triples) and parallel forms for a forced-choice assessment.

Reflections on First Year of ATA Use in an Established Testing Program
Jon Lehrfeld, Educational Testing Service

We discuss our solutions to some practical problems we experienced when introducing ATA into an established statewide K-12 testing program. Problems range from psychometric to computational to communication with content experts. We also look forward to problems we foresee arising in transitioning from year one to year two of ATA.

Practical Implications of Automated Test Assembly for Large-scale Multi-Stage Testing
Hanwook Yoo, Educational Testing Service; Jianbin Fu, Educational Testing Service

This study aims to deliver practical tips to implement the sophisticated ATA procedures based on experience in large-scale multistage testing of admission assessment. We will review the efficiency of partitioning the item pool and discuss the importance of testing security-related attributes, along with content and statistical specifications.

Automated Test Assembly Application for Multistage Testing of an English Language Assessment
Tongyun Li, Educational Testing Service; Shuhong Li, ETS; Venessa Manna, Educational Testing Service; Lixiong Gu, Educational Testing Service

This study is an empirical application of automated test assembly (ATA) for multistage adaptive testing of a newly-developed English assessment. The ATA implementation is based on mixed integer programming. The results provide implications regarding the use of ATA to assemble test forms that meet psychometric/content requirements under practical constraints.

An Automated Item Pool Assembly Framework for Maximizing Item Utilization for CAT
Hwanggyu Lim, Graduate Management Admission Council; Kyung (Chris) Han, Graduate Management Admission Council

This study introduces a new pool assembly framework called Honeycomb automated pool assembly (HAPA), which is developed to build parallel item pools of Graduate Management Admission Test. In computerized adaptive tests, HAPA will be exponentially efficient in producing a massive number of pools and make the maintenance of pools straightforward.

Discussant:
David Shin, **Pearson**

138. Foundational Competencies in Educational Measurement: NCME Task Force Consensus and Debate

Organized Discussion
8:00 to 9:30 am
Marriott: Floor 5th - Chicago Ballroom E

What are "foundational competencies in educational measurement"? What knowledge, skills, and abilities must modern students of educational measurement possess in order to continue learning and growing in our field? In October of 2021, NCME President Derek Briggs charged a 12-member Task Force to "develop and maintain foundational competencies in educational measurement." A year later, the Task Force engaged NCME membership in discussion of a draft report presenting three competency domains and five subdomains, as well as examples of how educational measurement careers and curricula develop these competencies. In this symposium, Task Force members will present their final report on Foundational Competencies in Educational Measurement, including a description and justification for each domain and subdomain in their framework. Three discussants who were not members of the Task Force will provide commentary on this report: 1) Is the framework coherent, defensible, and useful? 2) Are any foundational competencies missing, superfluous, or unambitious? 3) How can or should the field and NCME continue to support consensus around foundational competencies in educational measurement? This symposium debates the Task Force's foundational competencies on conceptual grounds. A complementary, preceding symposium debates how foundational competencies develop in measurement programs and manifest in careers.

Session Organizers:
Derek Christian Briggs, **University of Colorado Boulder**
Andrew Ho, **Harvard Graduate School of Education**

Presenters:
Deborah Bandalos, **James Madison University**
Matthew James Madison, **University of Georgia**
Michael C. Rodriguez, **University of Minnesota**
Michael Russell, **Boston College**
Stefanie A. Wind, **University of Alabama**

Discussants:
Ying Cheng, **University of Notre Dame**
Laura Hamilton, **American Institutes for Research**
David Torres Irribarra, **Pontificia Universidad Católica de Chile**

139. Innovative Methodologies in Computational Statistics

Coordinated Paper Session
8:00 to 9:30 am
Marriott: Floor 5th - Chicago Ballroom F

The world is rapidly changing thanks to the massive amounts of available data and computing power. Psychometrics and educational measurement are changing in tandem: rich data on test takers collected from different sources are becoming available in addition to traditional scoring: from keystrokes and response times to real-time video feed. Emerging testing setups range from remote testing to game-based collaborative learning. Larger data sets create large-scale, high-dimensional computational problems that require new methodologies: fast algorithms, data structures, and software implementations that generate results and meaningful insights in reasonable time and exploit parallel and cloud resources. This has been an active research area in other domains, yet in psychometrics, statistical inferences such as parameter estimation and hypothesis testing still largely rely on decades-old methodologies. This session highlights new computational methodologies for making statistical inferences in psychometrics: multigrid methods for estimating high-dimensional latent variable models; robust inference methods for IRT models; and distributed statistical inference for latent variable models. These innovations will enable richer user models, making sound inferences that are robust to data contamination such as cheating and disengagement; and reimagining ways for collecting, storing and analyzing test data efficiently and securely.

Session Organizer:
Oren Livne, **Educational Testing Service**

Participants:

Fast Multigrid Algorithms for High-Dimensional Maximum Likelihood
Oren Livne, Educational Testing Service

High-dimensional latent variable estimation via expectation-maximization requires a prohibitively expensive integral evaluation due to the curse of dimensionality. An alternative is gradient decent, which however converges slowly. Based on multigrid solvers' success in other scientific and engineering fields, we develop a fast multigrid algorithm for maximizing a non-linear likelihood functional.

Minimax Robust Inference in IRT Models
Michael Fauss, ETS

Robust estimation is proposed as a principled approach to dealing with recent challenges in educational testing. The problem of minimax robust estimation of a person parameter under incomplete knowledge of the item response function is introduced, and examples of the corresponding estimators for two types of uncertainty are discussed.

Distributed Statistical Inference for Latent Variable Models
Xiang Liu, Educational Testing Service

In this talk I will present a class of distributed statistical inference methods for IRT models. Important statistical properties of these methods will be discussed both theoretically and empirically through simulations. Finally, a real data set will be analyzed to demonstrate the utilities of the methods.

Learning from Computer Vision: "Seeing" Examinee Ability with a Convolutional Neural Network Model
Ikkyu Choi

In this talk, we will introduce a convolutional neural network model for estimating examinee ability under the unidimensional item response theory. The model makes a connection between data arising from educational measurement and images, opening up possibilities to utilize computing methods that have fueled great advances in computer vision.

Discussant:
Matthew Johnson, **ETS**

140. Use of Metrics and Thresholds in AI Scoring Model Evaluation

Coordinated Paper Session
8:00 to 9:30 am
Marriott: Floor 5th - Denver/Houston

The automated scoring of constructed responses is an area of psychometrics undergoing rapid development. There are new use contexts added each day, and new use contexts challenge the "standard practices" for how to build and evaluate automated scoring models. Given the wide variety of engines, statistical models, item types, score scales, and stakes, model evaluation practices should be reconsidered, specifically statistical metrics and thresholds. This session presents five papers which address the use of model evaluation metrics in AI scoring. The first paper surveys the literature to capture the current landscape of AI model evaluation. The second paper examines the quadratic weighted kappa and highlights its sensitivities. The third paper focuses on the proportional reduction in mean squared error and discusses its advantages and limitations. The fourth paper uses empirical data to assess the current industry guidelines under different use contexts, engines, score scales, etc. The fifth paper demonstrates how the PRMSE can be used during test development to select mixtures of AI-scored items with varying levels of psychometric quality. Together, these papers start a new conversation about which metrics psychometricians should be using to evaluate automated scoring models and how they should be used in practice.

Session Organizer:
Jodi Casabianca-Marshall, **Educational Testing Service**

Chair:
Sharon Slater, **ETS**

Participants:

How to Evaluate Your Automated Scoring Engine: Theory and Reality
Magdalen Beiting-Parrish, CUNY Graduate Center

QWK: An "Unreliable" Measure of Interrater Agreement?
Jennifer L. Lewis, University of Massachusetts Amherst; Jodi Casabianca, ETS

On the Proportional Reduction in Mean Squared Error for AI Model Evaluation
Jodi Casabianca-Marshall, Educational Testing Service; Daniel F. McCaffrey, ETS

Do the Current Model Evaluation Guidelines Work? Lessons Learned From Real Data
Ourania Rotou, New Meridian; Sharon Slater, ETS

Impact of the Quality of AI-Scored Items on Mixed-Format Test Score Reliability
Seyma N. Yildirim-Erbasli, Concordia University of Edmonton; Jodi Casabianca, ETS; Ourania Rotou, New Meridian; Sharon Slater, ETS

Discussant:
Susan Lottridge, **Cambium Assessment, Inc**

141. [SIGIMIE Session] Advancing Perspectives on Practice Analysis for Credentialing Examinations

Organized Discussion

8:00 to 9:30 am

Marriott: Floor 5th - Los Angeles/Miami

The process for determining the exam content in certification and licensure testing tends to differ from educational achievement testing in that the content on educational tests are often dictated by state or national standards, while certification and licensure organizations must develop and provide validity evidence for exam content specific to their domain. Unfortunately, the Standards for Educational and Psychological Testing provides little guidance on the conduct of practice analysis for credentialing organizations. Traditional methods for practice analysis have become so ingrained that research on methods for practice analysis over the past 20 years has been sparse. This session is designed to facilitate a discussion around the philosophy and methodologies for conducting a practice analysis for credentialing. Therefore, we have invited panelists from a variety of backgrounds to discuss not only their own explorations into methods for practice analysis, but also to discuss potential future directions and challenges. Specific topics for discussion include the creation of individualized test blueprints based on topics relevant to a person's specific practice; use of external data sources to provide validity evidence to support blueprint design; the unique challenges and solutions from small credentialing organizations; and competency models for practice analysis, among others.

Session Organizer:
Michael R Peabody, **National Association of Boards of Pharmacy**

Moderator:
Robert Thomas Furter, **Physician Assistant Education Association**

Presenters:
Andrew Dwyer, **American Board of Pediatrics**
Brett P. Foley, **Alpine Testing Solutions**
Pamela Kaliski, **ABIM**
Patricia Muenzen, **ACT**

142. [Joint Session with AERA Division D] State of the Field: Gender and Racial Equity in Educational Measurement

Organized Discussion

9:50 to 11:20 am

Marriott: Floor 5th - Chicago Ballroom B/C

National employment reports show that while the demographic makeup of the workforce is now trending more female and racially/ethnically diverse than ever before, workplace inequities still persist in influential and well-compensated positions. The past three decades of employment reports within the educational measurement field mirrors these findings, making us question, what factors are holding women back from achieving parity with men? To address this question, Women in Measurement, AERA, and NCME have partnered to produce a first-of-its-kind study on workplace equity in the educational measurement community. Our study is the first to examine intersecting marginalized identity groups (e.g., women of color) and the field's perceptions of employment diversity, equity, and inclusion (DEI) practices. In this session, we will present our preliminary findings of a census survey administered to students and professionals affiliated with WIM, AERA, and NCME on key indicators including social identity, employment position, educational training, professional experiences, salary and perceptions of DEI.

Session Organizer:
Ye Tong, **National Board of Medical Examiners**

Presenters:
Thao Vo, **Washington State University**
Susan Lyons, **Lyons Assessment Consulting**
Ye Tong, **National Board of Medical Examiners**
Felice Levine, **American Educational Research Association**
Nathan Bell, **American Educational Research Association**

143. Cheating Detection Using Machine Learning and Deep Learning Methods

Coordinated Paper Session
9:50 to 11:20 am
Marriott: Floor 5th - Chicago Ballroom D

The recent boost of online testing with remote proctoring provides new opportunities and challenges to measurement practices. One challenge in online testing is test security while one opportunity is the richness of assessment data that can be collected in online testing. Essentially, both structured and unstructured product and process assessment data can be collected during test administration, such as item responses, item response time, and clickstreams. It is expected that the new data types collected in online testing can facilitate cheating detection or aberrant response detection. Given the nature of such data from multiple sources, new methodologies are called for in cheating detection. In recent years, different machine learning methods have been explored for cheating detection. However, some areas still remain untapped. This session intends to demonstrate how different machine learning and deep learning algorithms can be used for cheating detection using item responses, item response time, clickstream data, and augmented data from psychometric analysis and machine learning methods. Further, how cheating detection can be conducted using natural language processing for tests consisting of constructed-response text data in addition to tests consisting of dichotomous items.

Session Organizer:
Hong Jiao, **University of Maryland**

Chair:
Manqian Liao, **Duolingo**

Participants:
An Autoencoder-Based Algorithm for Checking the Existence of Item Preknowledge in Computer-Based Testing
Yiqin Pan, University of Florida
Predicting Normal and Aberrant Behaviors based on Sequence Modeling of Test-Takers Clickstreams using LSTM, RNN, and N-Gram
Steven Tang, eMetric; Zhen Li, eMetric LLC
Integrating Psychometric Analysis and Machine Learning to Augment Data for Cheating Detection
Hong Jiao, University of Maryland; Guiyu Li, East China Normal University; Todd Zhou, Univeristy of Maryland; Shudong Wang, NWEA
Plagiarism Detection Using Human-in-the-loop AI
Manqian Liao, Duolingo, Inc.; Sinon Tan, Duolingo; Basim Baig, Duolingo
Machine Learning Algorithms for Detecting Answer Similarity in Open Ended Responses
William Skorupski, Data Recognition Corporation

Discussant:
Gregory Cizek, **University of North Carolina at Chapel Hill**

144. Holistic Admissions with Test-Optional Policies: Application Essays, Recommendation Letters, and Other Factors

Coordinated Paper Session
9:50 to 11:20 am
Marriott: Floor 5th - Chicago Ballroom E

Despite the increasing popularity of holistic admissions approach, it is unclear how this operationalized and implemented, especially under the test-optional policy. Further, the value and fairness issues associated with the use of qualitative information, such as application essays and recommendation letters, may require research and explorations to improve the field's understanding and provide evidence regarding the effectiveness and fairness of such uses. In this session, we will address related questions from multiple angles: 1) Racial and gender differences in the language/content of 2.5 million teacher recommendation letters from CommonApp; 2) How admissions officers practice holistic admissions, especially in relation to diversity and equity considerations based on a survey and interviews with 126 admission officers through NAGAP; 3) The quality and content of 38,000 application essays from a large university and their relationship with admissions decisions and first-semester GPA; 4) Comparing different methods of analyzing application essays and their relationship with test scores and socioeconomic background based on 60,000 University of California applicants. Together, these four papers would provide new ideas and information to further the discussion and exploration of a valid and fair admissions process for the U.S. higher education.

Session Organizer:
Guangming Ling, **Educational Testing Service**

Chair:
Ou Lydia Liu, **ETS**

Participants:

What's in a Letter? Using Natural Language Processing to Investigate Systematic Differences in Teacher Letters of Recommendation
Brian H Kim, CommonApp Inc.

As the first study of its kind, we examined the racial and gender-related differences in the language/content of 2.5 million teacher recommendation letters from CommonApp. There exist salient linguistic differences in letters across gender, but less evidence for differences across race – except in the case of highly competitive admissions.

Holistic Admissions in Graduate Schools: A Survey of Admissions Officers and Students
Sara Haviland, ETS; Joseph Paris, West Chester University; Reginald M Gooch, Educational Testing Services

Based on responses to surveys and interviews provided by admissions officers and prospective graduate students, we examined how the holistic admissions approach is implemented in practice, especially in relation to diversity and equity considerations.

What is the Value of Application Essays? An Exploration of its Role in Shaping College Admissions Decisions and Post-enrollment Outcomes
Sugene Cho-Baker, ETS; Brent Bridgeman, ETS; Michael Flor, Educational Testing Service; Guangming Ling, Educational Testing Service

We evaluated the quality and content of 38,000 application essays from a large university using multiple natural language process tools and automated writing evaluation tools, and revealed that the essay quality and content had complex relationships with outcomes such as admission status and post-admission success indicators.

Weighing Algorithmic Tradeoffs: Observations from 60,000 Admission Essays
Klint Kanopka; David Lang, Stanford University; A J Alvero, Stanford University

We compared different methods of analyzing application essays, including a few latest computational methods, and examined their relationship with test scores and socioeconomic background based on 60,000 University of California applicants. Our results suggest the latest computational tools may not always be the most appropriate initial path forward.

Discussant:
Li Cai, **UCLA**

145. Analytics and Design Considerations to Inform Test Development

Paper Session
9:50 to 11:20 am
Marriott: Floor 5th - Chicago Ballroom F

Chair:
Xaviera Gonzalez-Wegener, **UCL Institute of Education**

Participants:

Examining Response Time and Examinee Performance Under a Logistic Regression Modeling Framework
Seongeun Kim, University of North Carolina at Greensboro; Yang Zhao, American Board of Internal Medicine

Using real data from a medical certification exam, this study employs logistic regression models to investigate the time examinees take to respond to individual items and their exam performance, along with item characteristics and examinee demographics. The results are especially meaningful for test development, psychometrics, and medical education.

Psychometric and Design Considerations in Early Educational Assessment
Tony Albano, University of California, Davis; Robin Hojnoski, Lehigh University; Kristen Missall, University of Washington; David Purpura, Purdue University; Xiaochen Xu, University of California, Davis

In this presentation, we explore how familiar psychometric methods and test development practices from K12 education do and do not apply to the unique context of early education. Examples come from the LLAMA project, where we are creating adaptive touch-screen math assessments for preschoolers.

The Use of Complex-Structure Items in Multi-stage Testing
Paulius Satkus, Graduate Management Admission Council; Christine DeMars, James Madison University

Multi-stage tests (MSTs) may measure multiple constructs using simple or complex structure items or a combination, which we compared in a simulation. Bias and RMSE did not greatly vary across item structure. The potential difficulties of developing complex-structure items may not be worth the small benefits of administering simple-structure items.

Developing a State-of-the-Art Universal Screener for K-8 Mathematics
Bozhidar M. Bashkov, IXL Learning; Yao Xiong, Roblox Corporation; Christina Schonberg, IXL Learning; Luke Corazza, IXL Learning; Kate Mattison,
IXL Learning

Using evidence-centered design (Mislevy et al., 2003) and the latest developments in measurement, such as maximum priority index (Cheng & Chang, 2009) and embedded standard setting (Lewis & Cook, 2020), we developed an efficient yet powerful CAT universal screener for K-8 mathematics to classify students relative to standard-based achievement levels and grade.

Discussant:
Jonathan Rubright, **National Board of Medical Examiners**

146. The Digital SAT: the Impact of Changes

Coordinated Paper Session
9:50 to 11:20 am
Marriott: Floor 5th - Chicago Ballroom G/H

In 2022 the College Board announced that the SAT will be taken digitally beginning in 2023 for international students and 2024 for U.S.- based students. The digital SAT is shorter than the current SAT. Calculators will be allowed on the entire math section and shorter reading passages that cover a wider range of topics will be used for reading and writing sections. The SAT Suite will continue to measure the knowledge and skills that students learn in high school and that matter most for college and career readiness. Also, the digital SAT will still be scored on a 1600 scale and scores can be used to track growth across the SAT Suite of Assessments over time. Several studies have been conducted to evaluate the impact of the changes made to the digital SAT. In this session, five presentations will focus on the introduction of the key features of the digital SAT and studies on item quality and efficiency of reading and writing test, shortened timing of SAT tests, psychometric targets of the digital SAT, and linking for the current and digital SAT. Analyses results will be shared to address the issues brought by the changes made to the digital SAT.

Session Organizer:
Weiwei Cui, **College Board**

Participants:
Introduction of the New Digital SAT
Thomas Proctor, College Board
Digital SAT Reading and Writing Pilot to Examine Item Quality and Efficiency
Ying Lu, College Board; Judit Antal, College Board
The Psychometric Target for Assembling Digital SAT *Siang Chee Chuah, College Board; Oliver Zhang, College Board; Wei S. Schneider, College Board*
Timing Study for Digital SAT
Wei S. Schneider, College Board; Weiwei Cui, College Board; Denny Way, College Board; Sunhee Kim, College Board; Thomas Proctor, College Board
Linking Analyses for the Current and Digital SAT
Tim Moses, College Board

Discussant:
Richard Melvin Luecht, **University of North Carolina at Greensboro**

147. Research Blitz: On Various Topics from Test Design and Scale Validation to Modeling of Response Bias and Missing Data

Research Blitz Session
9:50 to 11:20 am
Marriott: Floor 5th - Denver/Houston

Chair:
Mubeshera Tufail

Participants:
K-Fold Cross-Validation for Factor Analysis
Kyle Nickodem, University of North Carolina - Chapel Hill; Peter Halpin, UNC-Chapel Hill; Carly Tubbs Dolan, New York University

Drawing upon machine learning literature, we propose a principled approach to assessing scale dimensionality via exploratory and confirmatory factor analysis utilizing k-fold cross-validation. Replication of a scale structure across the k subsamples (i.e., folds) demonstrates structure stability while model performance is aggregated across folds to inform scale development decisions.

Development and Validation of an Operationalisable Model of Critical Thinking
Sundance Zhihong Sun, The University of Melbourne; Zhonghua Zhang, University of Melbourne; Bruce Beswick, The University of Melbourne; Sandra Milligan, The University of Melbourne

The literature on critical thinking reveals difficulties in operationalising the widely adopted framework developed by Facione (1990). The study reported here was undertaken to develop a more operationalisable model of critical thinking. The new model was validated through the design and use of assessment instruments among students from different cultures.

Detecting Response Bias in Rating Scales with an Interaction Map
Jinwen Luo, UCLA; Minjeong Jeon, UCLA

We present a new approach to detecting response bias using an interaction map approach. Dependencies between respondents and item response categories, unexplained by the person and item parameters, are visualized as respondent-category distances in the map, revealing the presence, patterns, and size of potential response bias in the data.

Leveraging Public Occupational Data in Evidence-Centered Adult Assessment Design
Brendan Longe, University of Massachusetts Amherst; Maria Elena Oliveri, University of Nebraska Lincoln; Kevin P. O'Rourke

This study analyzes data from the federally sponsored Occupational Network (O*NET) database to inform the design of assessments for adult learners and workers based on a comprehensive domain analysis used to identify workplace relevant competencies including work Skills, Contexts, and Activities required for jobs in zones 1-3.

Using Expert Raters in a Validity Study for Diagnostic Modeling
Madeline Schellman; Laine Bradshaw, Pearson; Hollylynne Lee, North Carolina State University; Shannon Clark, University of Georgia

To examine the external validity of a diagnostic concept inventory (DCI) measuring middle grades students' misconceptions in probabilistic reasoning, we investigated the agreement consistency of model-based diagnoses and expert judgements of student classifications. Results provide evidence to inform the validity argument for using DCI results to identify instructional intervention needs.

Bayesian Exploratory Rating Scale Model for the Mixture of Response Styles
Denis Federiakin, Johannes Gutenberg University Mainz; Andreas Maur, Johannes Gutenberg University of Mainz; Lisa Martin de los Santos Kleinz, Johannes Gutenberg University of Mainz; Marie-Theres Nagel, Johannes Gutenberg University of Mainz

This paper presents the General Mixture Response Styles Model that models item response styles on Likert scales by placing equality constraints on item difficulties across latent classes. To make the parameter interpretation tractable, we use strong priors according to the classification of response styles developed for the model.

Missing Data in College Surveys: A Monte Carlo Simulation Study
Shimon Sarraf, Indiana University; Dubravka Svetina Valdivia, Indiana University

We investigate person parameter recovery with the graded response model by applying missing data handling techniques to empirical higher education survey data. Specifically, we ask how well listwise deletion, Amelia II, and MICE-CART can recover person parameters under various missing proportions and mechanisms (MCAR, MAR, NMAR).

A Comparison of Thurstonian IRT Model and Triplet-2PL Model
Jianbin Fu, Educational Testing Service; Xuan (Adele) Tan, ETS

In practice the Thurstonian IRT model is commonly used to calibrate forced-choice items with three statements (triplets) although it works directly with two statements. The recently developed Triplet-2PL model calibrates triplets directly. The current study compares both models on simulated and real data and supports the Triplet-2PL model for triplets.

148. Test Security Breaches: Prevalence, Detection Strategies, and Decision Making

Coordinated Paper Session
9:50 to 11:20 am
Marriott: Floor 5th - Los Angeles/Miami

While test security has been a major concern of testing programs for many years, it is arguably now a heightened concern for many programs due to shifts in testing administration procedures that have occurred since the pandemic. This session explores several topics relevant to this new era of test security. The session begins with a presentation examining indications of test security violations across a wide range of programs, as well as an in-depth look at security flags for a program that shifted from exclusively test center administration to a mix of test center and online proctoring administration. The second presentation introduces a computationally efficient method of calculating answer similarity statistics, enabling testing programs to more quickly identify and respond to security threats. The third presentation introduces an item type that is intended to help in the detection of item preknowledge, providing testing programs with a content strategy to enhance test security. The final presentation introduces a framework, based on Bayesian decision theory, which testing programs can use to arrive at a decision on whether some examinees committed test fraud. The session will include commentary from an expert in statistical detection of test fraud.

Session Organizer:
Carol Eckerly, **Educational Testing Service**

Participants:

Examination of Test Security Results for an Online Proctored Program
Kirk Becker, Pearson; Jinghua Liu, Pearson

Efficient Answer Similarity Analysis Using Dichotomous Scores
Carol Eckerly, Educational Testing Service; Ben Babcock, Association for Materials Protection and Performance; Kylie Gorney, University of Wisconsin-Madison; Mridul Aanjaneya, Rutgers University

Designed for Detection: Chameleon Clones and their Power to Detect Preknowledge
Sarah Linnea Toton, Caveon Test Security

Use of Bayesian Decision Theory in the Detection of Test Fraud
Sandip Sinharay, Educational Testing Service; Matthew Johnson, ETS

Discussant:
James Wollack, **University of Wisconsin**

149. Test Comparability Around the World: Methodological Challenges and Solutions

Coordinated Paper Session
11:40 to 1:10 pm
Marriott: Floor 5th - Chicago Ballroom B/C

In this coordinated session, we will discuss challenges and proposed solutions to comparability of tests from around the world, from higher-education admission tests in China and Sweden as well as global language tests, and to tests for schools: non-cognitive tests in Chile, and cognitive tests in The Netherlands. Specifically, the first paper presented in this symposium describes the methodological approach to evaluating the comparability of three mathematics tests of Gaokao, the college admission tests in China, the national test, and two designed by a local province. The second paper presents challenges of comparing multidimensional non-cognitive tests on school climate in Chile. The third paper discusses sampling considerations for building a concordance between the Duolingo English Test and other language tests. The fourth paper discusses methodological challenges to compare tests that assess the readiness of Dutch end-of-primary school children. Last but not least, the last paper discusses the effects of the revisions of the Swedish SAT, especially on the impact of the anchor design on comparability of different versions of the test.

Session Organizer:
Stella Kim, **University of North Carolina at Charlotte**

Chair:
Alina A. von Davier, **Duolingo**

Participants:

Comparability of Three Mathematics Tests for College Admission in China
Stella Kim, University of North Carolina at Charlotte; Chunlian Jiang, University of Macau; Chuang Wang, University of Macau; Jincai Wang, Suzhou University

This study attempts to examine the equivalence of three mathematics tests of Gaokao (college entrance exams of China) and illustrate the methodological procedure to establish concordance using a single group design with an item response theory (IRT) approach.

The Comparability of School Climate Measures at Schools in Chile
Jorge González, Pontificia Universidad Católica de Chile

In this work, we make use of a multidimensional graded response model to analyze a School Climate scale of the Chilean QEC, and explore different approaches of test equating to establish the comparability of these measures.

Sampling Considerations for Building Concordance Tables
Ramsey Cardwell, Duolingo; Steven Nydick, Duolingo; J.R. Lockwood, Duolingo; Alina A. von Davier, Duolingo

We present a case study of building concordance tables between scores from the Duolingo English Test (DET) and other English language proficiency tests. We discuss implementation challenges, including estimating the distribution of reporting error, dealing with multiple test records for the same person, and standard error estimation.

Comparability of Dutch End-of-primary School Tests
Marieke van Onna, Cito

This study investigated the comparability of Dutch end-of-primary school tests. Though no significant DIF was detected, comparing the results with a control background variable indicated some level of bias. Thus, the talk will discuss alternative comparability methods.

Test Revisions and Comparability: The Swedish SAT
Marie Wiberg, Umeå University; Inga Laukaityte, Umeå University

In this presentation the focus is on the discussion of the current process for the revision of the anchor test in the SweSAT. One of the focuses of this revision is on how to construct the anchor test to make it fair for as many test takers as possible.

Discussant:
Tim Moses, **College Board**

150. New Approaches to Some Contemporary Problems in Evaluating Achievement and Growth

Coordinated Paper Session
11:40 to 1:10 pm
Marriott: Floor 5th - Chicago Ballroom D

Standardized achievement tests have played an essential role in US schools since their widespread adoption during WWI. Achievement testing has accelerated with the digital revolution ever since and most of the well-known issues to deploying test scores to evaluate student performance have been resolved. In recent years however, research is also increasingly focused on the remaining challenges to the successful use of testing outcomes to measure, evaluate, and predict achievement growth. This session presents several new approaches to selected contemporary topics in the evaluation and prediction of achievement growth. Topics in growth measurement include the definition and construction of growth measures and, by implication, the needed elaboration on their associated scales and norms, and the setting of acceptable or adequate standards for growth. For growth evaluation settings in the presence of experimental or natural intervention, a perspective that treats points of learning recovery of students and schools as unknown and conditionally random is presented. In prediction, a new strategy that makes more comprehensive use of students' academic test score histories for predicting college and career readiness for students and the use of the Kalman Filter to make efficient and conditionally unbiased predictions that leverages prior test score histories are explored.

Session Organizer:
Yeow Meng Thum, **NWEA**

Chair:
Yeow Meng Thum, **NWEA**

Participants:
Developing Scales and Norms for Growth: Goals, Constraints and Consequences
Yeow Meng Thum, NWEA
Unlocking the Black Box: Understanding Adequate Growth Percentiles through Standard Regression Analogs
Katherine Furgol Castellano, Educational Testing Service; Daniel McCaffrey, Educational Testing Service; Joseph A. Martineau, Educational Testing Service
Capturing How Soon Student Learning Bounces Back in the Post-pandemic Era
Yong Luo, NWEA
Integrating Multiple Assessment Data in Developing College and Career Readiness Measures
Hong Jiao, University of Maryland; Jinglei Ren, University of Maryland; Robert Lissitz, University of Maryland
Improved Inference in Periodic Testing with State-space Models and Growth Estimates
Thomas Christie, NWEA; Garron Gianopulos, NWEA; Carson Cook, NWEA; Yeow Meng Thum, NWEA

Discussant:
Stephen Sireci, **University of Massachusetts Amherst**

151. [SIGIMIE Session] Big Data, Big Change, Big Decision

Organized Discussion
11:40 to 1:10 pm
Marriott: Floor 5th - Chicago Ballroom E

Big data is becoming a transformative tool for all facets of education. How to use big data, how to adapt to big changes and how to make big decisions in measurement are critical questions to be discussed and clarified. The NCME Big Data in Educational Measurement SIGIMIE, aligned with the upcoming 2023 NCME Annual Conference, proposes an organized discussion consisting of a panel of 4-5 speakers who are experts of Big Data in relation to educational measurement and related areas. The structure of the panel session will include discussion sections for all speakers, crafted questions from SIGIMIE members, and interactive collection of questions from real-time attendees. This panel session will highlight conversations about how to prepare for integration of data science and psychometrics in measurement. Specifically, discussion questions may include: what kind of skills and knowledge are in urgent needs, how could we better disassemble the product of this new integration, how big data could be better aligned with the needs of education, curriculum and policy. The panel will be invited representing from five important and practical perspectives: (1) testing and assessment industry, (2) university and training programs, (3) high-tech companies, (4) journal editorial board, and (5) education policy.

Session Organizers:
Qiwei He, **Educational Testing Service**
Tiago A. Caliço, **Independent**

Moderators:
Qiwei He, Educational Testing Service
Tiago A. Caliço, Independent

Presenters:
Kadriye Ercikan, Educational Testing Service
Steven Culpepper, University of Illinois at Urbana-Champaign
Chun Wang, University of Washington
John Whitmer, Institute of Education Statistics
Yao Xiong, Roblox Corporation

152. Validating a Writing Trait Model for Formative Use

Coordinated Paper Session
11:40 to 1:10 pm
Marriott: Floor 5th - Chicago Ballroom F

Efforts have been made to develop trait scoring systems for automated writing evaluation (AWE). However, these systems can suffer from an over-reliance on human judgment, which tends to be highly correlated even when they are asked to rate student essays on distinct scales. An alternative approach is to develop a multidimensional model of writing that posits traits that represent measurable dimensions of variation in student essays, measured using natural language processing (NLP) features designed to capture aspects of writing quality, style, and other constructs of interest to the writing teacher. Such a model can be trained by conducting confirmatory factor analysis on a large corpus of student writing. This session presents a series of papers that demonstrate progress in validating a multidimensional writing trait model for formative use. In particular, we show that one such model has value for predicting scores over and above the holistic scores produced by a traditional automated essay scoring (AES) system, that it has strong test-retest reliability, and that it can be used to describe the relationships between different types of writing products, such as stylistic differences among essays written to different tasks, or the relation between student plans and their final written texts.

Session Organizer:
Paul Deane, ETS

Moderator:
Tenaha O'Reilly, Educational Testing Service

Participants:
Using a Writing Trait Model to Understand Teacher Scoring and Feedback
Paul Deane, ETS; Matthew Myers
Test-Retest Reliability of Writing Traits in Classroom Data
Paul Deane, ETS; Duanli Yan, ETS; Chen Li, ETS
Exploring the Relationship between Planning Features and Essay Performance
Yi Song, Educational Testing Service; Paul Deane, ETS; Chen Li, ETS
Exploring Trait Differences Between Genres in Two Corpora
Duanli Yan, ETS; Paul Deane, ETS; Chunyi Ruan, ETS

Discussant:
Danielle McNamara, Arizona State University

153. Vendor Collaboration That Supports State Solutions

Organized Discussion
11:40 to 1:10 pm
Marriott: Floor 5th - Chicago Ballroom G/H

Reimagining Assessments did not stop at recommendations for improved assessments, policies, and systems. It also invited an opportunity to improve internal processes and the vendor collaborations that shape them. This session dives into the work of State Education Agencies (SEAs) and the facets and responsibilities of federal and state-mandated assessment programs. Some of the nation's largest and leading SEAs and testing vendors engage in a focused discussion that assesses our current environment, as we work past competition and towards genuine cross-vendor collaboration and innovations. Panelists will explore successful and unsuccessful SEA experiences with vendor partnerships and share best practices. Vendor panelists respond to SEAs call for support by tackling peer review, operating, and meeting federal guidelines, ideal state characteristics, and collaborations that could improve them all, absent individual strategic advantages. They are committed to confronting and challenging internal company issues, attitudes, communication, transparency, and leadership approaches that negatively impact relationships and contract goals. Vendors are equally committed to collaborating with each other to create, share, and modernize best practices that include flexibilities that anticipate future growth and change. This session promotes clear communication and partnership that advances mutual goals, maximizes assessment community efforts, and improves our services to students.

Session Organizer:
Elda Garcia, **National Association of Testing Professionals**

Moderator:
Mary Anne Arcilla, **Educational Testing Service**

Presenters:
Andrew J. Middlestead, **Michigan Department of Education**
Zachary Warner, **New York State Education Department**
Vince Verges, **Florida Department of Education**
Chloe Torres, **NWEA**
Mark Johnson, **Cognia, Inc.**

154. Research Blitz: Test Scaling, Linking, and Equating

Research Blitz Session
11:40 to 1:10 pm
Marriott: Floor 5th - Denver/Houston

Chair:
Justin L. Kern, **University of Illinois at Urbana-Champaign**

Participants:

The Impact of Compromised Anchor and Non-Anchor Items on Equating Results
Siyu Wan, ABIM; Aaron Myers, American Board of Internal Medicine
Compromised anchor items can affect the equating process and undermine the validity of score interpretations. Little is known about the effect of compromised non-anchor items on equating. We evaluated the effect of compromised anchor and non-anchor items on scaling coefficients and equated scores via a simulation. Practical implications are discussed.

An Investigation of Context Effects on Equating for a Non-speeded Test
Linette P. Ross, NBME; Chunyan Liu, National Board of Medical Examiners
The impact of context effects on parameter estimation, score accuracy, and equating has been a psychometric concern. The purpose of this study is to evaluate if scrambled item blocks impact test performance and Rasch equating results. Study results will help practitioners identify context effects and their impact on equating.

The Impact of Different Drift Detection Methods on Rasch Scale Stability
Sarah Alahmadi, James Madison University; Andrew Jones, American Board of Surgery; Carol L Barry, American Board of Surgery; Beatriz Ibanez Moreno, American Board of Surgery
A group of well-established and newly introduced drift detection approaches were compared using simulated and operational data. Two methods exhibited superior performance in examinee ability recovery and classification accuracy in large samples. All methods performed similarly unfavorably in small samples. Failure rates varied when methods were applied to operational tests.

Evaluation of Flagging Criteria for Anchor Item Stability Analysis
Katherine Nolan, Curriculum Associates; Nina Deng, Kaplan INC.
Testing programs commonly use the anchor design to place student scores from different administrations onto one scale. D-squared and robust z are often used to ensure the anchor set is performing similarly. This study evaluates different criteria of d-squared and robust z for selecting the optimal anchor set under Rasch.

Revisiting Angrist et al (2021): Merging International Assessment Data Is Not Easy
Radhika Kapoor; Klint Kanopka; Ben Domingue, Stanford University
Can data from different international assessments (e.g. PISA, TIMSS, SACMEQ) be linked into one informative score? This paper examines the requirements of this kind of linking endeavor. We utilize simulation studies to show how (and under which conditions) this linking might not be robust for country comparisons and rank ordering.

Exploring Linking Scores between Modes with Common Items Non-Equivalent Groups
Shichao Wang, ACT
Mode effects should be studied when an exam moves from paper-based testing to computer-based testing. A new mode linking method was proposed and intensively evaluated in the current study. The results show that the proposed method may provide beneficial information when adjusting scores for mode differences for certain test subjects.

Jackknife Methods to Evaluate Item Response Theory Linking and Equating Stability
Sunhee Kim; Denny Way, College Board
This paper proposes Jackknife methods for evaluating an item that impacts the stability of linking and equating results. The inferential and empirical evaluation criteria are also discussed. Thru simulations and empirical data, we observed the benefits of those statistics on evaluation of equated score stability.

Accumulation of Systematic Error in Linear Equating Chains
Benjamin Andrews, Inteleos
Sinharay and Holland (2010) demonstrated methodologies to evaluate equating methods in ways that do not give an advantage to any single method. Similar methods are extended to the linear equating case. The effects of violations of assumptions on accumulated equating error in a chain of equatings are also investigated.

Comparing Simultaneous and Separate Calibration Linking Methods
Guangyun Liu; Won-Chan Lee, University of Iowa; Hyung Jin Kim, University of Iowa; YoungKoung Kim, College Board
This simulation study evaluates the performance of simultaneous linking for item parameter recovery, in comparison to linking through separate calibration. Multiple linking designs are considered to examine the amount of accumulated linking error over time.

155. Modeling Test Taking Behaviors Through Process Data

Paper Session
11:40 to 1:10 pm
Marriott: Floor 5th - Los Angeles/Miami

Chair:
Kristin M. Morrison, **Curriculum Associates**

Participants:

Comparing and Validating Process Data Indicators of Test-Taking Effort on Constructed-Response Items
Militsa Ivanova, University of Cyprus; Michalis Michaelides

The study aimed to construct, compare, and validate process-data indicators of test-taking effort on constructed-response items in PISA. Results suggested that item response time or a combination of response time and number of actions may serve as better indicators of engagement than the number of actions performed on an item.

"Unused Time" and "Mastery" Indicators for Examining Test-taking Behaviors
Elena Papanastasiou, University of Nicosia; Michalis Michaelides; Joseph A. Rios, University of Minnesota

The purpose of this study is to describe and evaluate two novel indicators of test-taking behaviors that utilize a combination of response and timing data to better understand and describe test-taking effort in ILSAs. These indices will be empirically estimated with data from the fourth-grade e-TIMSS 2019 mathematics assessment.

Priming Examinees to Give Good Effort: Differential Utility across Student Groups
Katarina Schaefer, James Madison University; Sara Finney, James Madison University; Mara McFadden, James Madison University

Answering questions about intended effort prior to completing a test resulted in higher examinee self-reported effort and response-time effort for first-year college students. Moreover, fewer examinees were filtered from the dataset due to low effort. However, for upper-class students, only females were impacted by this strategy to promote good effort.

Test-Taking Behaviors on an Admissions Test: Variations by Race and Gender
Alaysia Marie Brown, Educational Testing Service (ETS); Sugene Cho-Baker, ETS; Guangming Ling, Educational Testing Service

The current study explored response process patterns on a higher education admissions test using keystroke data to examine whether associations between test-taking behaviors and performance outcomes differed across demographic groups. We found that examinees' test-taking behaviors, and associations between test-taking behaviors and performance outcomes, varied by race/ethnicity and gender.

Modeling Students' Item Revisit Behavior on TIMSS 2019 Math Items
Jihang Chen, Boston College; Zhushan Mandy Li, Boston College

Identifying examinees' test-taking strategies is of increasing interest in the educational measurement field. We apply a speed-accuracy-revisit model to TIMSS process data to understand the relationship between examinees' speed, ability, and revisiting tendency. Results show a weak relationship between examinees' revisiting tendency and ability, and different test-taking strategies across subgroups.

Discussant:
Anne Traynor, **Purdue University**

156. [Joint Session with AERA Division D] Revision of the Standards for Educational and Psychological Testing

Organized Discussion
1:30 to 2:30 pm
Marriott: Floor 5th - Chicago Ballroom B/C

Under the auspices of the AERA-APA-NCME Standards Management Committee, the three organizations have begun taking the first steps in revising the 2014 edition of the Standards for Educational and Psychological Testing. The Standards reflect the three organization's shard guidance on testing on such issues as validity, reliability, and fairness in testing, and reflect the highest ideals and expectations regarding test development, administration, use, and interpretation. Launching a new edition is a project that demands the expertise and input of diverse experts and stakeholders. This session held under the auspices of the Management Committee is one essential step in clarifying the scope, focus, and issues essential for consideration in the next edition of the Standards. Attendees are encouraged to bring their questions, concerns, and wisdom to this joint AERA-NCME session.

Session Organizer:
Kristen Huff, **Curriculum Associates**

Moderator:
Kristen Huff, **Curriculum Associates**

Presenters:
Michael C. Rodriguez, **University of Minnesota**
Doris Zahner, **CAE**
Cara Cahalan Laitusis, **ETS**
Rochelle Michel, **Smarter Balanced**
Guillermo Solano-Flores, **Stanford University**

157. Successful NLP Approaches to Automate Scoring of NAEP's Reading Assessment

Coordinated Paper Session
1:30 to 2:30 pm
Marriott: Floor 5th - Chicago Ballroom D

In 2021, the National Center for Education Statistics (NCES) issued a challenge to psychometricians and data scientists: Apply natural language processing (NLP) to replicate human scoring on students' responses to open-ended reading items on the National Assessment of Education Progress (NAEP). Almost two dozen teams submitted entries, and NCES selected six winning submissions. Each one used innovative methodologies, such as deep learning, large language models, and neural networks. These winning submissions developed automated scoring models that substantially improved upon the human-to-automated scoring accuracy of previous essay-based research. The average degradation in accuracy from machine-scored to human-scored was -0.028 QWK or less, which is strikingly accurate, especially considering the 0.905 QWK agreement among human scorers. In this session, NCES will discuss the potential role of automated scoring for NAEP. Three Presenters will explain their approaches and results, including their potential for use across large-scale assessments. Discussants will share its program of research on applying NLP to NAEP assessment items and how the agency will use the research in scoring for NAEP reading assessments, more recent developments in this research, evaluate the use of automated scoring in other subjects (e.g., mathematics), and discuss how they will address bias in automated scoring.

Session Organizer:
Eunice Greer, **Department of Education**

Participants:
Automated Scoring for Reading Comprehension via In-context BERT Tuning
Andrew Lan, University of Massachusetts at Amherst
Automated Scoring of Reading Items Using Computationally-Efficient Transformer Models
Susan Lottridge, Cambium Assessment, Inc
Combining Linguistic Features with Deep Neural Network Models to Fine-Tune Response Predictions for NAEP Reading Items
Arianto Wibowo, Measurement Incorporated

Discussants:
Peggy Carr, **National Center for Education Statistics**
John Whitmer, **Institute of Education Statistics**

158. Through-year Assessment Systems: Impacts on Educational Decision Making

Coordinated Paper Session
1:30 to 2:30 pm
Marriott: Floor 5th - Chicago Ballroom E

As the conference theme suggests, the field of K-12 educational measurement is being challenged to improve the information provided to educators so they can make better educational decisions. Rising to this challenge, states are pursuing innovative assessment systems that meet federal accountability requirements while also providing actionable information for instructional decision making throughout the year. Given these initiatives, we describe the types of through-year models that have emerged, including but not limited to models that have emerged through the U.S. Department of Education Innovative Assessment Demonstration Authority. In this session, we (a) share motivations for seeking through-year assessment models from both the district and state leader perspectives, (b) synthesize the opportunities these innovations offer to inform better decision making, as well as the practical and measurement challenges that arise, (c) delineate the distinctions between through-year model types from a measurement perspective, and (d) facilitate discussions among the panel of Presenters—that includes district leaders, state leaders, measurement professionals, and through-year assessment developers—to share insights from designing and implementing these innovations to help inform the future of through-year assessment systems and K-12 educational measurement.

Session Organizer:
Laine Bradshaw, **Pearson**

Moderator:
Trent Workman, **Pearson**

Participants:
Through-year Assessment: District-level Decision Making
Peter Leonard, Chicago Public Schools
Through-year Assessment Models: Measurement and Validity Considerations
Amy Reilly, Pearson; Laine Bradshaw, Pearson; Melinda Montgomery, Pearson
Through-year Assessment: State-level Decision Making
Chris Rozunick, TEA

159. Investigating Measurement Invariance in Noncognitive Assessment

Paper Session
1:30 to 2:30 pm
Marriott: Floor 5th - Chicago Ballroom F

Chair:
Yoav Bergner

Participants:
Investigating Measurement Invariance in NAEP Student Questionnaire Index Items
Yichi Zhang; Young Yee Kim, American Institutes for Research; Xiaying Zheng, American Institutes for Research

In order to make valid group comparisons of questionnaire indices, measurement invariance (i.e., items functioning identically) needs to be established across groups. This study applies the measurement invariance testing procedure developed for polytomous items and the Region of Measurement Equivalence framework to evaluate the practical significance of measurement noninvariance.

Investigation of Measurement Invariance in Cross-Cultural Research for the Affective Domain
Cigdem Toptas; George Engelhard, UGA

The purpose of this study is to examine the invariance of several affective scales used in international educational research. In addition to examining item fit, this study stresses the description of methods for evaluating person fit **for** affective scales. Data from PISA (2012) are used in this study.

Discussant:
Jonathan Weeks, **Educational Testing Service**

160. Fairness and Equity in Assessment

Paper Session
1:30 to 2:30 pm
Marriott: Floor 5th - Chicago Ballroom G/H

Chair:
Jung Yeon Park, **George Mason University**

Participants:
Equity-minded Assessment: A Framework for Born Socio-culturally Responsive Assessment
Edynn Sato, Sato Education Consulting LLC

This framework, based on research from multiple disciplines, addresses the following: How can an assessment be designed to build on students' cultural knowledge and their sociocultural contexts in order to promote equity? How can such an assessment provide comparable data and evidence for a given inference across diverse learners?

Exploring Fairness Perspectives on 11+ Selection in the Caribbean
Jerome De Lisle, University of the West Indies; Stephen Geofroy, The University of the West Indies; Murella Sambucharan-Mohammed, The University of the West Indies; Carla Kronberg, UWI St Augustine; Tracey Michelle Lucas, University of the West Indies; Sharon Phillip, The University of the West Indies; Nalini Ramsawak-Jodha, University of the West Indies, St. Augustine, Trinidad; Nisha Harry, The University of the West Indies; kristy Phillip, The University of the West Indies

We constructed an expanded framework of fairness philosophical perspectives that includes the modern fairness/justice viewpoints proposed by Sandel, Sen, and ul-Haq. We use this framework to analyse stakeholder viewpoints on various policies for early test-based selection at 11+ taken from a series of public consultations in Trinidad and Tobago

Academic Genealogy as a Methodological Approach Towards Anti-racist Measurement Research
Jade Caines Lee, University of Kansas
Identifying and examining mentoring relationships can be a powerful methodological lever in uncovering the trajectory of measurement ideologies. Known as academic genealogy, measurement practitioners and scholars can begin to understand and examine how ideological legacies in educational measurement are created, preserved, and embedded in broader systems of power.

Discussant:
Fiona Hinds, **Independent Consultant**

161. Demonstrations: Session 2

Demonstration Session
1:30 to 2:30 pm
Marriott: Floor 5th - Denver/Houston

Chair:
Kevin Krost, **Virginia Polytechnic Institute and State University**

Participants:

Detecting Compromised Items using an R package
Chansoon Lee, American Board of Internal Medicine
This study will demonstrate an R package: 1) determining thresholds for sequential procedures to detect compromised items; 2) calculating a series of statistics based on classical test theory and item response theory; 3) detecting compromised items using a hybrid threshold approach; 4) plotting a series of statistics of flagged items.

Operation REQUISITE: Re-envisioning Educational Quantitative User Information: Shiny Interface to Explore
Eric M Ho; Brittany Boyd, American Institutes for Research; Juanita Hicks, American Institutes For Research; Cadelle Hemphill, AIR
There is demand from teachers and practitioners to re-envision the reporting of finer-grained educational data. We propose a framework that provides guidance on choosing the most useful visualizations to make better sense of quantitative data. We introduce a Shiny application using this framework to guide users and create these visualizations.

SmartItems TM: using Automatic Item Generation for Test Security
Sergio Araneda, University of Massachusetts Amherst; David Foster, Caveon Test Security
We will present the concept of “SmartItem”, developed at Caveon Test Security. We will explain what Smart-items are, how to create them in Caveon’s platform Scorpion, and what types of items can be created using this technology.

Using Video Survey Administration (ViSA) as a Reliable Tool for Data Collection
Christopher Martin Amissah, Morgan State University; Taj Rollins; Alaa Alkhalaiwi, Morgan State University; R. Trent Haines, Morgan State University
ViSA is a new computer-based method for collecting survey data in educational and health-related research. It is an interactive technique that allows participants to enter responses to a prerecorded video presentation of surveys via touchscreen or mouse clicks. Pilot data supported ViSA as a reliable alternative to text-based computer survey.

162. Tools and Perspectives on Assessment Literacy

Paper Session
1:30 to 2:30 pm
Marriott: Floor 5th - Los Angeles/Miami

Chair:
Dukjae Lee, **University of Massachusetts Amherst**

Participants:

Developing a Tool for Educational Leaders to Select High-quality Assessment Literacy Initiatives
Yelisey A Shapovalov; Carla M. Evans, National Center for the Improvement of Educational Assessment
Many educators need additional professional learning related to assessment. We developed a screening tool that helps states or districts select a high-quality assessment literacy program for K-12 educators using a systematic literature review and expert feedback. Key domains include content, implementation features, and institutional readiness.

Teacher Learning and Collaboration: Increasing Data Informed Decision Making in the Classroom
Tara Kintz, Michigan Assessment Consortium; John Lane
While there is increasing focus on the formative assessment process, less is known about efforts to support teacher learning and collaboration that leads to increased data informed decision making in the classroom. This session explores a statewide initiative and data collected over a three-year period to increase educator assessment literacy.

Assessment Literacy through the Lens of Leadership and Validity.
Ian Hembry, MetaMetrics
Assessment literacy is lauded as a cost-effective measure to improve schools. Yet, a lack of understanding often results in unintended uses being introduced. This study explores how leadership principles can facilitate assessment literacy and bolster a tests validity argument through an online course as an exemplar.

Discussant:
Thanos Patelis, **Johns Hopkins U & University of Kansas**

163. Improving Measures of Opportunity to Learn (OTL) to Address Systemic Inequity

Coordinated Paper Session
2:50 to 4:20 pm
Marriott: Floor 5th - Chicago Ballroom B/C

To leverage measurement for better decisions, we need to contextualize achievement data and provide more nuanced, complete, and actionable analyses and reporting. This session highlights measurement of contextual variables, including opportunity to learn (OTL) and income, to draw attention to systemic inequities that contribute to differences in achievement. These five papers use data from several sources including NAEP, the RAND American Teacher Panel, the National Longitudinal Study of Adolescent to Adult Health, the U.S. Census, the American Community Survey, as well as Title I funding data from one U.S. state to demonstrate approaches to measuring aspects of OTL and other contextual variables across levels of educational systems with the purpose of supporting educational equity.

Session Organizer:
Leslie Nabors Olah, **Educational Testing Service**

Chair:
Leslie Nabors Olah, **Educational Testing Service**

Participants:
(In-)equitable Distribution of School Opportunities to Learn Mathematics
Leslie Nabors Olah, Educational Testing Service; Carolina Lopera-Oquendo, CUNY Graduate Center
Informing Equity Through Indicators of Social, Emotional, and Civic Learning Opportunities
Margarita Olivera Aguilar, Educational Testing Service; Laura Hamilton, American Institutes for Research; Sam Rikoon; Olasumbo Oluwalana, **Educational** *Testing Service; Jennifer Lentini, Educational Testing Service (ETS)*
Investigating the Relationship Between School Culture, OTL, and Students' Longer Outcomes
Constance Lindsay, University of North Carolina at Chapel Hill
Evaluating the Effect of Title I Funds on Resource Equity: A Case Study
Elizabeth A. Fernandez-Vina, New Jersey Department of Education
Measuring Socioeconomic Segregation in 10 Dimensions
Josh Leung-Gagné, Stanford University

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

164. [SIGIMIE Session] How Can Statewide Accountability Testing Improve Student Learning?

Organized Discussion
2:50 to 4:20 pm
Marriott: Floor 5th - Chicago Ballroom D

This session is an extension of last year's symposium "Reimagining Assessments: The Responsibility is Ours!" For that session, the NCME State and Local Assessment Leaders SIGIMIE assembled a diverse panel committed to collaborations that support assessment systems that support learning for all students. For NCME 2023, a new and diverse panel comprising state assessment leaders, university professors, educational policy practitioners, and testing vendors will discuss the innovations with which they have been involved to use assessments to support students. They will share their goals and discuss additional work that needs to be done to increase the value of assessments by leveraging technology, simplifying policies and processes, diversifying assessment partners, and pursuing vendor collaboration. Panelists will explore and discuss how assessments can be better designed to support and inform teachers as they work to meet the needs of all students. They will provide updates on their post-pandemic innovations, including progress monitoring and through-year testing. They will share what went well, what did not, and directions for the future. Accountability demands, complexities, and challenges will be woven into the discussion. Panelists will interact with the audience to confront negative perceptions, misinformation, data misuse, political challenges, and ways to meet ESSA requirements.

Session Organizer:
Zachary Warner, **New York State Education Department**

Moderator:
Elda Garcia, **National Association of Testing Professionals**

Presenters:
Trent Workman, **Pearson**
Andrew J. Middlestead, **Michigan Department of Education**
Stephen G Sireci, **University of Massachusetts, Amherst**
Darin Kelberlau, **Millard Public Schools**
Christina Laster, **National Parents Union**

165. [SIGIMIE Session] Towards Culturally Relevant Assessment: Reconceiving How to Incorporate Culture Into Teaching Measurement

Organized Discussion
2:50 to 4:20 pm
Marriott: Floor 5th - Chicago Ballroom E

The field of educational measurement has come under scrutiny for its insensitivity to cultural factors. Some have argued that a failure to account for such factors has resulted in anti-testing sentiment amongst minoritized populations, which has ultimately mitigated the perceived credibility and utility of many educational assessments. This session, hosted by the Educators of Measurement Special Interest Group, proposes an engaging, organized discussion around incorporating cultural context into teaching educational measurement by intentionally emphasizing demonstrations of teaching by leaders in the field. Specifically, the session consists primarily of two teaching demonstrations that respectively focus on cultural and community validity and quantitative critical race theory. Each demonstration includes an introductory lecture on the presented topic as well as planned activities that require audience participation. Following each demonstration, attendees will have the opportunity to ask demonstrators topical- and pedagogical-related questions. The session will conclude with a summation of potential solutions raised by participants for better incorporating cultural context into teaching as well as a contemplative/discussion period for attendees to consider the types of changes that they can employ in their teaching to better address societal concerns around cultural insensitivity in educational measurement.

Session Organizers:
Joseph A. Rios, **University of Minnesota**
Anne Corinne Huggins-Manley, **University of Florida**
Brian French, **Washington State University**

Presenters:
Pohai Kukea Shultz, **University of Hawaii at Manoa**
Kerry Englert, **Seneca Consulting, LLC**
Michael Russell, **Boston College**

166. Simulating Large-Scale Assessment Data: Tools and Practice

Coordinated Paper Session
2:50 to 4:20 pm
Marriott: Floor 5th - Chicago Ballroom F

This paper session showcases three R packages for simulating large-scale data with inherent complex survey designs including item responses, survey responses, weights, survey attributes, and plausible values. Implications will be discussed, and an example of NAEP-like synthetic data will be presented with simulation methodology included.

Session Organizer:
Ting Zhang, **American Institutes for Research**

Chair:
Ting Zhang, **American Institutes for Research**

Participants:
lsasim: An R Package for Simulating Large-Scale Assessment Data
Yuan-Ling Liaw; Leslie Rutkowski, Indiana University; David Rutkowski, Indiana University
Dire: An R Package for Latent Regressions and Plausible Values Generation
Paul Bailey, American Institutes for Research; Ting Zhang, American Institutes for Research; Emmanuel Sikali
Simulating Large-Scale Assessment Data Using the R Package simsem
Alexander M Schoemann, East Carolina University; Terrence D Jorgensen, University of Amsterdam
A Use Case of Simulating NAEP-like Data
Sinan Yavuz, American Institutes for Research; Paul Bailey, American Institutes for Research; Ting Zhang, American Institutes for Research; Emmanuel Sikali

Discussant:
Leslie Rutkowski, **Indiana University**

167. Tackling Through-Year Assessment Topics from a Practitioner's Point of View

Coordinated Paper Session
2:50 to 4:20 pm
Marriott: Floor 5th - Chicago Ballroom G/H

Due to the increasing demand and popularity of the through-year (TY) assessment, testing companies were given the opportunity to operationally implement them, and to use the empirical data to reveal some of the TY assessment mysteries. In this symposium, we try to tackle some TY topics from a practitioners' point of view. The first two papers present several TY assessment designs and shed the pros and cons through empirical and simulated data. The third and fourth papers use empirical data collected from a special field-test design that enables the investigation of item parameters calibrated from the fall, winter, and spring administrations to explore the season effect on item calibration and the possibility of using a longitudinal multidimensional Item Response Theory model and a latent growth item response model to track the longitudinal growth of students' ability and investigate the item parameter estimation accuracy. The last paper introduces three approaches in utilizing student ability estimates from the previous administrations to improve the performance of next administrations.

Session Organizer:
David Shin, **Pearson**

Participants:
Measuring Growth in a Through-year Assessment
Melinda Montgomery, Pearson; Kuo-Feng Chang
Interim Assessments with Cumulative Blueprint Design
Jie Li, NCS - Pearson; Jeffrey Hauger, University of Massachusetts; Richard O'Neill, Peoria Unified School District
Season Effect on the Field-Tested Item Calibration
Tianshu Pan, Pearson; John Vito Binzak, Pearson Clinical Assessment; David Shin, Pearson
Modeling Learning Growth using Longitudinal Data with Multiple Measures from Adaptive Assessment
Yanyan Tan; David Shin, Pearson
Borrowing Theta from Previous Administrations to Next administrations in Through-Year Assessment
Yawei Shen; Yang Lu, Pearson

Discussant:
William A. Lorie, **Center for Assessment**

168. [SIGIMIE Session] Harmonize Tradition and Innovation: Scaling, Linking, and Equating in Technology-Enhanced Measurement

Organized Discussion
2:50 to 4:20 pm
Marriott: Floor 5th - Denver/Houston

Technology has profoundly changed our traditional ways of teaching, learning, and testing over the past decades. This session brings together professionals who work closely with technology-enhanced (TE) assessments and frameworks in various testing fields to share their expertise in scaling, linking, and equating (SLE) practices. An educational measurement scholar will discuss these practices and provide insights regarding SLE methods for the present and future TE measurement.

Session Organizer:
Mengyao Zhang, **National Conference of Bar Examiners**

Moderator:
Kyung Yong Kim, **University of North Carolina Greensboro**

Presenters:
Xia Mao, **NBOME**
Larissa Smith, **NBOME**
Shu-chuan Kao, **NCSBN**
Jennifer Davis, **Amazon Web Services (AWS)**

Discussant:
Alina A von Davier, **Duolingo**

169. Computer Adaptive Testing: Models and Estimation

Paper Session
2:50 to 4:20 pm
Marriott: Floor 5th - Los Angeles/Miami

Chair:
Brendan Longe, **University of Massachusetts Amherst**

Participants:

Domain Ability Estimation in Shadow Test Assembly with Bifactor Models
Sangdon Lim, University of Texas at Austin; Seung W. Choi, University of Texas at Austin

This study examines the performance of domain ability estimation methods when items measure a bifactor structure, in the context of multidimensional shadow-test approach to computerized adaptive testing. The impact of different approaches for defining the scalar information for shadow-test assembly are examined, including directional information, Kullback-Liebler information, and mutual information.

Application of Constrained Confirmatory Mixture IRT Model in Computerized Adaptive Test
Minho Lee, University of California Los Angeles; Meredith Langi, NWEA

Although it is important to consider population heterogeneity in the measurement process, previous studies and common practice in CAT often assume the population being tested is homogeneous. This study investigates the applicability of constrained confirmatory mixture IRT model on CAT data, drawn from unidimensional IRT model with homogeneous population assumption.

Adaptive Design to Facilitate Alternate Ability Estimators in Group-score Assessments
Jiaying Xiao, University of Washington; Paul Adrian Jewsbury, Educational Testing Service; Usama Ali, Educational Testing Service; Peter van Rijn, ETS Global

The study was designed to explore the design features of adaptive testing to facilitate alternative ability estimators for group-level inference without drawbacks of plausible values. Five ability estimators were compared under different conditions. Results demonstrated with adequate adaptive designs, EAP-LRM and WLE may serve as a replacement to PVs.

The Effects of Uncontrolled Speededness on IRT Item and Ability Estimation
Jaime Malatesta, Graduate Management Admission Council; Paulius Satkus, Graduate Management Admission Council; Kyung (Chris) T. Han, Graduate
Management Admission

Many testing programs have transitioned from using Classical Test Theory (CTT) as their measurement framework to now using Item Response Theory (IRT). However, exam delivery frameworks, including time constraints, still reflect test-level CTT principles. This paper examines how this discrepancy between measurement and delivery frameworks can impact IRT estimation.

Dynamic, On-the-Fly, and Hybrid Multistage Testing with the Bi-Factor Model
Shawna Goodrich, Department of National Defence; Okan Bulut, University of Alberta

Several novel multistage testing approaches were extended from a unidimensional to a bifactor structure with real data. Then, a Monte Carlo simulation study was conducted to compare test and module length conditions. Longer tests and shorter modules in the earlier stages and longer modules at later stages produced better results.

Discussant:
Kirk Becker, **Pearson**

170. The Impact of Pandemic on Testing Industry

Coordinated Paper Session
4:40 to 6:10 pm
Marriott: Floor 5th - Chicago Ballroom B/C

This session examines how the pandemic has disrupted testing programs and changed the outlook of the testing industry across a variety of testing fields. All five papers in this session have common interests: how testing volume and test taker performance have changed prior to and after the pandemic. In addition, each individual paper has its own focus. Two papers explore how pandemic has affected credentialing industry, with one paper depicting an overall picture across 100 credentialing exams spanning several industries, and other focusing on regulatory exams and conducting subgroup analysis. The third paper examines the impact of the COVID-19 disruption on the K-12 assessment from multiple perspectives, including trend interpretation and use, operational procedures, and customer requests & program design. The last two papers explore the pandemic impact on high-stake admission tests: SAT and GRE® General Test, respectively. The fourth study uses statistical adjustment procedures to account for changes that are due to different factors and compares SAT cohorts before and after COVID-19 via this approach. The fifth paper evaluates the composition shift in the GRE® General test taking population, and its impact on performance trends.

Session Organizers:
Jinghua Liu, **Pearson**
Kirk Becker, **Pearson**

Participants:
Candidate Performance in a Pandemic: Exploring Pass Rate Changes in Credentialing
Susan Davis-Becker, ACS Ventures, LLC; Timothy Muckle, NBCRNA; Pooja Shivraj, American Board of OBGYN
This study evaluated the differences in candidate volume and pass rates from 2019 to 2020 and then 2020 to 2021 from credentialing industry annual reports. These findings were compared across professional industries and test administration modes. This presentation will summarize the investigation, findings, and recommendations for future research.
How Has the Pandemic Affected US Regulatory Exams?
Jinghua Liu, Pearson; Kirk Becker, Pearson; Nabeel Qazi, Pearson
This paper examines candidates' volume and pass rate changes in the context of several multi-state regulatory exams since 2016, not only at the total group level, but also at the subgroup level such as gender, ethnicity and education.
K-12 Assessment & the Pandemic: Navigating the Impact from Multiple Perspectives
Jennifer Dunn, Pearson; Jennifer Beimers, Pearson; Mark Roebeck, Pearson; Steve Fitzpatrick, Pearson
This study aims to examine the impact of the COVID-19 disruption on the K-12 assessment. Impacts will be explored from three perspectives: (1) trend interpretation and use, (2) operational procedures, and (3) customer requests & program design through analyses of empirical data resulting from multiple state assessment programs.
Comparing SAT Cohorts Before and After the COVID-19 Pandemic
Tim Moses, College Board; YoungKoung Kim, College Board
The study explores the pandemic impact on a high-stake admission test: SAT. It uses statistical adjustment procedures to account for changes that are due to different factors, such as changes due to pandemic versus changes due to college admission policy, and compares SAT cohorts before and after COVID-19.
Pre- and Post-pandemic Population Shift for a Higher-Education Admissions Assessment
Xuan (Adele) Tan, ETS; Hanwook Yoo, Educational Testing Service; James Carroll, ETS
This study looks at the changing landscape of testing for one major higher-education admissions assessment, the GRE® General test. Besides testing volume and overall performance trends, the test taker population characteristics will be compared pre- and post-pandemic to evaluate test-taker composition shift to shed light on changes in performance trends

Discussant:
Victoria Locke, **Istation**

171. Remembering Ron: Reflections on a Career and a Legacy

Coordinated Paper Session
4:40 to 6:10 pm
Marriott: Floor 5th - Chicago Ballroom D

Ronald K. Hambleton was one of the most influential and productive psychometricians of the late 20th and early 21st centuries. He not only made many contributions to the field in diverse areas, he also mentored and instructed many people around the world on psychometric theory and assessment practices. He influenced national and international testing programs, testing companies, and many people within the field. In this symposium, we honor these accomplishments through presentations by five of his colleagues who will focus on five different areas in which Professor Hambleton made monumental contributions: item response theory, detecting differential item functioning, computerized-adaptive testing, cross-lingual assessment, and score reporting. These colleagues will discuss his work in these areas and also comment on Ron the person; that is, how his caring for others remains his most lasting impression. These presentations will be followed by discussant remarks from his friend from graduate school, with whom he worked with for over 40 years.

Session Organizer:
Stephen Sireci, **University of Massachusetts Amherst**

Chair:
Stephen Sireci, **University of Massachusetts Amherst**

Participants:
Item "Ronsponse" Theory: Professor Hambleton's Impact on IRT
Richard Melvin Luecht, University of North Carolina at Greensboro
Ronald K. Hambleton's Advances in Detecting Item Bias, Differential Item Functioning
Bruno D. Zumbo, University of British Columbia
Professor Hambleton's Contributions to Test Construction
Wim J van der Linden, University of Twente
Cross-lingual and International Assessment: The Hambleton Legacy
Jose Muniz, University of Oviedo; Jose Muniz, University of Nebrija
Improving More Than Tests: Ron's Influence on Score Reporting
April Zenisky, University of Massachusetts Amherst

Discussant:
Hariharan Swaminathan, **University of Connecticut**

172. Test Equity and Fairness from the Voices that Matter

Organized Discussion
4:40 to 6:10 pm
Marriott: Floor 5th - Chicago Ballroom E

Recovery from the pandemic has taken far longer than anyone expected. Exacerbated achievement gaps between historically marginalized students and their peers necessitates a clear call to action to create fair, equitable, and anti-racist assessments and assessment systems. This session will include the perspectives of minority, social justice, and parent advocates and experts who will share their recommendations for a reimagined assessment system. Our panelists will explore their definitions of equity and fairness, share past efforts, and provide potential solutions towards equitable assessments and assessment systems. We will facilitate a discussion about social justice challenges and anti-racist assessment solutions. Panelists will discuss the implications of unintended consequences under current assessment systems and the impact on various stakeholders, especially historically marginalized communities. Panelists will discuss current court cases that highlight state takeovers based on the current statewide assessment criteria. We will invite attendees to help us move beyond administering assessments for compliance and instead engage in discussions and efforts that increase the value of assessments and invite transformative accountability. With collective effort and a social justice intent, we have enormous potential to create fair and equitable solutions that could result in effective changes to current assessment system. It is time!

Session Organizer:
Elda Garcia, **National Association of Testing Professionals**

Presenters:
Jade Caines Lee, **University of Kansas**
Maria Armstrong, **Association of Latino Administrators and Superintendents**
Christina Laster, **National Parents Union**
Molly Faulkner-Bond, **WestEd**

173. Improving Teacher Decisions in the Mathematics Classroom Through Measurement

Coordinated Paper Session
4:40 to 6:10 pm
Marriott: Floor 5th - Chicago Ballroom F

This coordinated set of papers demonstrates how empirical evidence from an assessment development program in algebraic, statistical, and computational thinking helps teachers better interpret test results, provides useful information about student growth over time, and ties items and skills to score reporting and performance standard setting. This applied research leverages the BEAR Assessment System (BAS), which is designed to enhance the development and interpretation of student assessment information to guide better teaching decisions. The first paper describes how the National Math and Science Initiative has used the BAS along with associated software to allow the development of innovative assessment of college readiness skills in STEM; the second paper shows how a standard setting technique can be used to provide both validity evidence and helpful documentation for teachers, and the third and fourth papers describe and demonstrate a flexible technique for looking both at individual subscales and an overall composite dimension.

Session Organizer:
Karen Draney, **University of California, Berkeley**

Participants:
Developmental Levels are Embedded in a Construct Map
Yukie Toyama, University of California, Berkeley; Richard Brown, National Math and Science Initiative
Construct Mapping: Evidence of Internal Structure, and Interpretable Meaning for Teachers
Karen Draney, University of California, Berkeley; Richard Patz, UC Berkeley
Relationship Between the Composite and Subscales in Problem-Solving Using Mathematics
Perman Gochyyev, University of California, Berkeley; Mark Wilson, Berkeley School of Education, UC Berkeley
Changes from Pretest to Posttest for Assessment of Data-Based Decision-Making
Perman Gochyyev, University of California, Berkeley

Discussant:
Howard Everson, **CUNY Graduate Center**

174. Transforming K-12 Assessments: Providing Valid Data for Instructional Decisions, Equity, and Accountability

Organized Discussion

4:40 to 6:10 pm

Marriott: Floor 5th - Chicago Ballroom G/H

National Association of Assessment Directors (NAAD) will host a symposium with featured Presenters and a discussion around innovative ways that assessments can provide data that is meaningful for accountability and instruction especially as it relates to equity in the student population. The following topics will be explored by the Presenters; 1) Inequities that exist with the current system, 2) Create a system that is equitable, reliable, and timely, 3) Use of innovative assessments to support student learning, 4) Matching assessment use with intended purpose, 5) How to provide valid data to various stakeholders, 6) Interpretation of federal laws by states around accountability

Session Organizer:
Charlotte Gilbar, **Natrona County School District**

Presenters:
Annie Rae Clementz, **Illinois State Board of Education**
Edith Aurora Graf, **Educational Testing Service**
Horatio Blackman, **National Urban League**
Michael C. Rodriguez, **University of Minnesota**

175. [CODIT and AERA Division D EIC Invited Session] Recruitment and Retention of Minoritized Measurement Professionals

Coordinated Paper Session

4:40 to 6:10 pm

Marriott: Floor 5th - Denver/Houston

Recruitment and retention of minoritized measurement professionals is critical for the field of measurement to effectively represent multiple perspectives, drive the field for cultural responsiveness and address inequities in education. In this session we will (a) present findings from research studies focusing on the experiences of racially and ethnically minoritized women working in the field of educational measurement as well as (b) highlight systemic efforts to increase the representation of Black, Brown, and Indigenous (BBI) students into the field of educational measurement.

Session Organizer:
Kadriye Ercikan, **Educational Testing Service**

Participants:
"Young, Gifted, and Black": How Undergraduate Black Women Experience the Measurement Field
Jade Caines Lee, University of Kansas; Daniela Cardoza, The University of Iowa
A Survey of Recruitment and Retention Efforts
Joseph A. Rios, University of Minnesota; Jennifer Randall, University of Massachusetts
Career Opportunities for Women in Measurement in Latin America
Leslie Vanessa Rosales de Véliz, JML Measurement and Testing Services, LLC
Women in Educational Measurement: Overcoming the Challenge of Visibility
Njideka Gertrude Mbelede, Nnamdi Azikiwe University Awka

176. Gauging Student Understanding In-The-Moment Through the Formative Assessment Process

Coordinated Paper Session

4:40 to 6:10 pm

Marriott: Floor 5th - Los Angeles/Miami

In this series of descriptive papers, we explain how the Formative Assessment Process stimulates in-the-moment assessment of student performance through eliciting evidence of student understanding. Furthermore, the formative assessment process links eliciting evidence of student understanding to making in-the-moment instructional decisions based on this evidence. In this series of coordinated papers, we introduce the formative assessment process, explain how formative assessment cultivates evidence-based instructional decision making, provide a programmatic overview of how one state promotes the use of the formative assessment process, consider the evidence of formative assessment's impact on student learning, examine the use of formative assessment in one classroom, and discuss disciplinary-specific considerations of evidence-based decision making.

Session Organizer:
John Lane

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

Participants:

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

The first half of the paper outlines the principles of the formative assessment process. The second half of the paper details how the FAME program has operationalized these principles into five Components including Planning, Learning Target Use, Eliciting Evidence of Student Understanding, Feedback Use, and Instructional Decision Making.

Tightening the Links among Eliciting Evidence of Student Understanding, Feedback, and Decision Making in Classroom Assessment
Margaret Heritage

This paper explains how ambitious teaching connects to the formative assessment process and describes how formative assessment strengthens the feedback loop from providing clear learning targets, eliciting evidence of student understanding, providing feedback, and making instructional and learning decisions. Thus, the feedback loop is embedded in the formative assessment process.

How Effective Use of Formative Assessment Practices in the Disciplines can Tighten Eliciting Evidence of Student Understanding, Feedback, and Decision Making
Tara Kintz, Michigan Assessment Consortium; Amelia Gotwals, Michigan State University

This paper centers on the need for educators to learn about and more effectively use formative assessment practices deeply rooted in the specific disciplines in which they are implemented. This paper also examines how teachers deepen their knowledge in their specific discipline while also developing their formative assessment practice.

Understanding Eliciting Evidence of Student Understanding, Feedback, and Decision Making: Evidence from One Classroom
John Lane

This paper considers the extent to which a single teacher, working diligently to enact the formative assessment process, is able to use questioning to provide feedback and to make instructional decisions.

Discussant:
Edward Dean Roeber, **Michigan Assessment Consortium**

Participant Index

(Last Name, First Name, Session Number)

., Ketan, 091

Aanjaneya, Mridul, 148
Abad, Francisco J., 027
Abdalla, Widad, 099
Abebe, Rediet, 130
Abulela, Mohammed, 097
Abuosbeh, Zein, 097
Ackerman, Terry, 034, 039, 065-16
Adams, Elizabeth L., 129
Ahadi, Stephan, 013
Ahn, Dahwi, 057
Ahn, Jungwon Rachael, 116
Ahn, Sunyoung, 089-10
Akande, Christiana Aikenosi, 089-6
Alahmadi, Sarah, 062, 154
Alakbarova, Vafa, 091
Albano, Tony, 010, 060, 145
Ali, Usama, 056, 169
Alkhalaiwi, Alaa, 161
Allen, Jeff M., 090
Almond, Russell G, 043
Aloe, Ariel M., 110
Alonzo, Alicia, 067
Alvero, A J, 144
Amissah, Christopher Martin, 161
An, Lily, 057
An, Xiaozhu, 089-18
Andrade, Heidi L., 081
Andrade-Lotero, Luis Alejandro, 085
Andress, Tim, 074
Andrews, Benjamin, 154
Anghel, Ella, 103
Anguiano-Carrasco, Cristina, 030
Ankomah, Francis, 081
Antal, Judit, 089-16, 104, 146
Anyidoho, Abena, 028
Araneda, Sergio, 057, 095, 161
Arce Nazario, Rafael, 089-9
Arcilla, Mary Anne, 153
Armstrong, Maria, 172
Arslan, Burcu, 065-18
Arthur, Ann, 089-8
Asamoah, Nana Amma, 089-6
Avvisati, Francesco, 114
Awwal, Nafisa, 023
Ayers-Wright, Elizabeth, 094
Ayik, Bilgehan, 057

Babcock, Ben, 148
Badrinarayan, Aneesha, 071
Baghestani, Shireen, 014
Bahr, Jan Luca, 073, 081
Baidoo-Anu, David, 105
Baig, Basim, 143
Bailey, Paul, 006, 123, 166
Baker, Doris Luft, 074
Baker, Eva L., 088
Bakken, Sara, 130
Baldwin, Peter, 065-12
Ballard, Laura D, 089-7
Bandalos, Deborah, 138
Banerjee, Sube, 089-5
Bao, Yu, 092
Barragan Torres, Mariana, 089-18
Barry, Carol L, 154
Bartz, Kayla, 054
Basaraba, Deni, 028
Bashkov, Bozhidar M., 072, 089-18, 145
Bates, Meg, 089-18
Bay, Luz, 065-2
Becker, Kirk, 054, 083, 148, 169, 170
Becker, Rachael N, 129
Bediwy, Ahmed, 073, 094
Bei, Ni, 097
Beilstein, Shereen Oca, 089-18
Beimers, Jennifer, 170
Beiting-Parrish, Magdalen, 095, 116, 140
Bell, Nathan, 142
Belov, Dmitry I., 022
Belwalkar, Bharati, 029
Bennett, Randy, 071
Berenbon, Rebecca, 028
Bergner, Yoav, 105, 159
Berry, Yufeng, 016
Beswick, Bruce, 147
Betts, Joe, 089-2
Bezirhan, Ummugul, 103
Biancarosa, Gina, 052
Binici, Salih, 128
Binzak, John Vito, 167
Bishop, Kyoungwon Lee, 065-3
Blackman, Horatio, 174
Boatman, Angela, 057
Bolsinova, Maria, 114
Bolt, Daniel, 054, 087, 111
Bonifay, Wes, 100
Borgonovi, Francesca, 051
Borowiec, Katrina, 029, 057
Bowe, Anica, 112
Boyd, Aimee, 092
Boyd, Brittany, 161
Bradford, Lydia, 054
Bradshaw, Laine, 065-8, 074, 087, 147, 158
Bratsch-Hines, Mary, 057
Brennan, Robert, 089-10
Brice, Amanda, 046, 126
Bridgeman, Brent, 093, 144
Briggs, Derek Christian, 039, 068, 076, 100, 138
Britton, Tolani, 130
Brooks, Christopher, 110
Brown, Alaysia Marie, 155
Brown, Gavin T. L., 033
Brown, Naomi, 089-11
Brown, Richard, 173
Brunetti, Matthew, 090
Bryer, Jason, 081
Buchholz, Janine, 114
Buckendahl, Chad W., 017, 064
Buckley, Jack, 082, 124, 135
Bulut, Cigdem, 102
Bulut, Okan, 018, 022, 023, 029, 051, 063, 089-1, 096, 102, 108, 113, 124, 131, 169
Burchell, Diana, 097
Burgess, Yin, 097
Burgmanis, Girts, 097
Burke, Matthew, 083
Burkhardt, Amy, 065-5, 089-14, 131
Burstein, Jill, 130
Butterbaugh, Donna J, 097
Büyükkıdık, Serap, 080
Buzo-Casanova, Enrique, 086
Bynum, Bethany, 057
Cai, Li, 058, 069, 100, 144
Cai, Liuhan Sophie, 054, 116
Caliço, Tiago A., 151
Calliouet, Ruth, 120
Camara, Wayne J., 075, 121
Camargo Salamanca, Sandra Liliana, 057
Canbolat, Yusuf, 022
Cancado, Luciana, 065-17, 091, 104
Canto, Phil, 084
Cao, Canxi, 015
Cao, Yi, 106, 119
Cappaert, Kevin, 065-4, 091
Cardoso-Leite, Pedro, 113
Cardoza, Daniela, 175
Cardwell, Ramsey, 149
Carr, Peggy, 157
Carroll, James, 170
Carroll Miranda, Joseph, 089-9
Cartier, Salenah, 020
Casabianca-Marshall, Jodi, 140
Casas, Maritza, 022
Castaneda, Ruben, 037
Chalmers, R. Phillip, 058
Cham, Heining, 019
Cham, Heining, 023
Chan, Jason C.K., 057
Chang, Ammi, 057
Chang, Hua-Hua, 094
Chang, Kuo-Feng, 104, 167
Chang, Wanchen, 016, 065-19
Chao, Hsiu-Yi, 065-5, 089-2
Chavez, Carlos, 065-15
Chawla, Kamal, 089-16
Chen, Becky, 097
Chen, Cheng Te, 025
Chen, Dandan (Danielle), 081
Chen, Hong, 073
Chen, Jianshen, 011
Chen, Jie, 019
Chen, Jihang, 054, 106, 155
Chen, Jing, 060, 136
Chen, Jinsong, 027, 038
Chen, Jyun-Hong, 065-5, 089-2
Chen, Ofer, 105
Chen, Troy, 104
Chen, Xiaowen, 057, 073
Chen, Xing, 016
Chen, Yinghan, 132
Chen, Yunxiao, 089-2
Cheng, Yiling, 089-1
Cheng, Ying, 065-4, 077, 111, 113, 138
Chiang, Yi-Chen, 029, 034
Chien, Yuehmei, 117
Chiu, Chia-Yi, 027, 116
Chiu, Ming Ming, 024
Cho, Sun-Joo, 089-12, 102
Cho-Baker, Sugene, 075, 144, 155
Choe, Edison M., 077
Choi, Hunwon, 031
Choi, Hye-Jeong, 016
Choi, Ikkyu, 066, 078, 119, 139
Choi, Jaehwa, 021, 031
Choi, Jinah, 094, 057
Choi, Jinnie, 012
Choi, Kilchan, 042, 088

Participant Index

(Last Name, First Name, Session Number)

Choi, Seung W., 004, 169
Choi, Youn-Jeng, 031, 022, 089-10, 097
Christie, Thomas, 101, 110, 150
Christopherson, Sara, 125
Chuah, Siang Chee, 146
Chung, Greg, 088
Chung, Gregory, 042
Chung, Jinmin, 031
Cirlot, La'Shea, 116
Cisterna, Dante, 105
Cisterna-Alburquerque, Dante, 105, 124
Cizek, Gregory, 091, 143
Clark, Amy, 014, 092, 105
Clark, Shannon, 147
Clauser, Amanda, 063, 107
Clauser, Brian, 065-12, 085, 107
Clauser, Jerome, 090, 107
Clementz, Annie Rae, 174
Close, Catherine, 060, 079, 125
Cobb, Paul, 110
Coggeshall, Whitney Smiley, 090
Cohen, Allan, 054, 089-10, 102
Coles, Jessica, 011, 065-2
Collard, Jasmine, 065-4
Colvin, Kimberly, 081
Cook, Carson, 101, 110, 150
Cook, Robert J., 017, 136
Corazza, Luke, 145
Çorbacı, Ergün Cihat, 029
Corrado, Kelly, 088
Corrin, Linda, 033
Cox, Olivia, 076
Crawford, Brandon, 089-6
Crespo Cruz, Eduardo Javier, 091
Crinion, Miriam, 057
Croft, Michelle, 014
Cruce, Ty, 075
Cruz, Tania, 011, 065-2
Cuhadar, Ismail, 024
Cui, Mengyao, 099
Cui, Weiwei, 146
Cui, Wenju, 066
Cui, Ying, 012, 018, 124
Culpepper, Steven, 132, 151
Curnow, Christina, 029
Custer, Michael, 037

Dadey, Nathan, 120
Dai, Shenghai, 081
Daisher, Ted, 126
Dallas, Andrew, 086
Damico, Danielle, 028
Davis, Jennifer, 083, 168
Davis, Laurie, 076, 092, 094, 126
Davis-Becker, Susan, 106, 170
Davison, Mark, 052
Deane, Paul, 066, 152
De Boeck, Paul, 019, 080, 102
de la Torre, Jimmy, 024, 061
De Lisle, Jerome, 057, 160
DeMars, Christine, 129, 145
Demir, Cihan, 035
Demirkaya, Onur, 089-17
Denbleyker, Johnny, 034
Deng, Jiayi, 073, 089-13
Deng, Jiayi, 096
Deng, Nina, 154
Denissov, Serguei, 093
Denner, Madelynn, 113
Deribo, Tobias, 022
Devasia, Nimmi, 093
Devkota, Rashmi, 089-5
DeWeese, Joseph, 052
Dhakal, Khagendra Raj, 010
Diao, Hongyu, 037
Diao, Qi, 016, 093, 137
Dilek, Ismail, 058
DiStefano, Christine, 019
Do, Tai, 065-15
Doğuyurt, Mehmet Fatih, 029
Domingue, Ben, 154
Domingue, Ben, 058, 096
Dong, Dongsheng, 054
Dong, Yixiao, 089-18
Donoghue, John, 129
Dorsey, David, 135
Doupe, Malcolm B., 089-5
Drame, Elizabeth, 112
Draney, Karen, 116, 173
Dresher, Amy, 068
Duan, Qizhou, 077
Duan, Yinfei, 089-5
Dumas, Denis, 089-18
Dunbar, Stephen, 065-9, 089-12, 089-15
Dunn, Jennifer, 170
Durrence, Debbie, 039
Dwyer, Andrew, 141
Dwyer, Kevin, 136
Dymchuk, Emily, 089-5

Eacker, Halley, 011, 065-2
Ece, Berivan, 089-14
Eckerly, Carol, 109, 148
Edi, Daniel, 022, 037
Edwards, Douglas, 089-9
Edwards, Kelly, 069
Egan, Karla, 032
El Masri, Yasmine, 081
Embretson, Susan, 058
Emery, Matthey, 124
Engelhard, George, 102, 159
Englert, Kerry, 030, 165
Ercikan, Kadriye, 082, 135, 151, 175
Erdemir, Aysenur, 026
Ersan, Ozge, 052
Estabrooks, Carole A., 089-5
Evans, Carla M., 074, 162
Evans, Josiah, 070
Everson, Howard, 039, 173

Fager, Meghan, 061
Faiello, Matthew, 057
Falk, Carl, 065-11
Fan, Fen, 063, 094
Fan, Meng, 016
Fan, Tingting, 092
Fang, Yu, 070
Faulkner-Bond, Molly, 071, 079, 172
Fauss, Michael, 053, 065-19, 066, 139
Federiakin, Denis, 147
Feiler, Jake, 012
Feinberg, Rich, 005, 109
Feldberg, Zachary, 112
Felline, Cosimo, 088
Feng, Tianying, 042, 088
Feng, Zechu, 061
Fernandez-Vina, Elizabeth A., 163
Ferrara, Steve, 014, 060, 106, 135
Feuerstahler, Leah, 016, 102, 116
Fienberg, Michael, 128
Filonczuk, Audrey, 111
Fina, Anthony D., 065-14, 073, 110
Finch, Holmes, 035
Fincher, Melissa, 136
Finn, Bridgid, 065-18
Finney, Sara, 155
Firoozi, Tahereh, 011, 063
Fischbach, Antoine, 113
Fishtein, Daniel, 093
Fitzpatrick, Steve, 117, 170
Flor, Michael, 144
Flowers, Tavia, 097
Flynn, Kylie, 028
Fogle, Thomas, 065-12
Foley, Brett P., 141
Forte, Ellen, 046, 136
Foster, David, 101, 161
Foster, Natalie, 082
Freeman, Jason, 089-9
French, Brian, 035, 081, 165
Frey, Bruce, 087
Frey, Sharon, 013
Fu, Jianbin, 065-15, 093, 137, 147
Fu, Yanyan, 034, 063, 099
Fujimoto, Ken, 065-11, 089-11
Furgol Castellano, Katherine, 057, 068, 150
Furter, Robert Thomas, 023, 141

Gagnon-Bartch, Johann, 118
Galib, Linda, 089-11
Gamo, Sylvie, 065-7, 113
Gao, Furong, 034
Gao, Rui, 026
Gao, Yizhu, 012, 017, 051
Garcia, Elda, 153, 164, 172
García-Minjares, Manuel, 086
Ge, Yuan, 116
Geofroy, Stephen, 160
Gerasimova, Daria, 115
Gershon, Richard, 089-14
Gianopulos, Garron, 110, 150
Gianopulos, Garron, 090
Gierl, Mark J, 011, 063, 065-18, 113
Gijbels, Liesbeth, 131
Gilbar, Charlotte, 174
Gocer Sahin, Sakine, 097
Gochyyev, Perman, 116, 173
Godek, Ben, 136
Goh, Sugyung, 031
Goldhammer, Frank, 022
Gong, Brian, 057, 074
González, Jorge, 149
Gonzalez-Wegener, Xaviera, 121, 145
Gooch, Reginald M, 057, 144
Goodman, Joshua, 012, 017, 062, 086, 094
Goodrich, Shawna, 169
Goodwin, Amanda, 089-12
Gorgun, Guher, 018, 022, 063, 065-15, 096, 105, 113
Gorney, Kylie, 063, 109, 148

Participant Index

(Last Name, First Name, Session Number)

Gotwals, Amelia, 125, 176
Grabovsky, Irina, 065-12, 097
Graf, Edith Aurora, 015, 174
Green, Anson, 130
Greer, Eunice, 157
Greiff, Samuel, 051
Griger, Cassondra, 094
Griph, Gerald, 016
Grochowalski, Joseph, 123
Gu, Dai, 057
Gu, Lixiong, 137
Gugiu, Mihaiela Ristei, 097
Guo, Hongwen, 082, 089-3, 097, 102, 106, 119
Guo, Wenjing, 073
Guo, Yage, 089-15
Gwak, Yelin, 031, 097
Gyll, Sean, 097

Habermehl, Kyle, 085
Habing, Brian, 047, 089-3, 111
Hahnel, Carolin, 022
Haines, R. Trent, 161
Halpin, Peter, 012, 110, 147
Hamdani, Maria, 055
Hamilton, Laura, 068, 138, 163
Han, Kahee, 081
Han, Kyung (Chris) T., 034, 063, 077, 087, 137, 169
Han, Suhwa, 080
Han, Youngjin, 069
Han, Yuting, 026, 104
Han, Zhuangzhuang, 119
Hansen, Mark, 011
Hao, Jiangang, 041, 053, 066
Happel, Jay, 050
Harik, Polina, 078, 085
Haring, Samuel, 024
Harris, Deborah J, 045, 133
Harry, Nisha, 057, 160
Hathcoat, John D, 129
Hauger, Jeffrey, 167
Haviland, Sara, 144
He, Qiwei, 035, 048, 051, 097, 114, 151
He, Surina, 018, 029, 051
He, Yi, 065-6
He, Yong, 104, 111
Heffernan, Neil T, 118
Hellman, Scott, 085
Hembry, Ian, 162
Hemphill, Cadelle, 161
Hendrickson, Amy, 057
Henson, Robert, 122
Heritage, Margaret, 176
Hernandez, Diley, 089-9
Hernandez, Philip, 054
Hicks, Juanita, 080, 161
Hille, Kathryn, 119
Himelfarb, Igor, 065-11
Hinds, Fiona, 160
Hirt, Ashley, 105
Ho, Andrew, 039, 067, 138
Ho, Emily, 089-14, 099
Ho, Eric M, 015, 161
Ho, Jordan L, 065-17
Ho, TsungHan, 065-3
Hoben, Matthias, 089-5
Hodge, Kari, 026, 112
Hoffman, Alexander, 060
Höft, Lars, 073, 081
Hogan, Melissa, 130
Hojnoski, Robin, 145
Holtzman, Steven, 093
Hong, Hyeri, 010, 028, 034
Hong, Hyeri, 010
Hong, Minju, 078
Hong, Yuan, 013
Hoover, Jeffrey, 087, 116
Hopkins, David, 120
Hornung, Caroline, 065-7, 113
Howell, Heather, 085
Howell, Jessica, 075
Hu, Xiangen, 082
Hua, Cheng, 008, 027
Huang, Chun-Wei, 028
Huang, Mingya, 089-11
Huang, Qi, 054
Huang, Sijia, 089-4, 108
Hubert, Barbara, 130
Hudson, Kim, 032
Huff, Kristen, 030, 095, 124, 156
Huff, Stacy R, 095
Huggins-Manley, Anne Corinne, 057, 089-6, 111, 165
Huh, NooRee, 016
Huo, Huade, 123
Hwu, Bo Sien, 025

Ibanez Moreno, Beatriz, 154
Ihlenfeldt, Samuel Dale, 072
Ing, Marsha, 110
Ingrisone, James, 029
Ingrisone, Soo, 027, 029
Inostroza Fernández, Pamela Isabel, 065-7, 113
Ivanova, Militsa, 155

Jackson, Janine, 062, 085, 095
Jackson, Kara, 110
Jackson, Yvette Yvette, 073
Jafari, Amir, 085
Jansen, Thorben, 073, 081
Jeffries, Jay, 057
Jeon, Eunjeong, 022
Jeon, Minjeong, 015, 147
Jeong, Hyo, 056
Jeppson, Haley, 047
Jewsbury, Paul Adrian, 089-3, 169
Ji, Feng, 022
Ji, Xuejun Ryan, 018
Jia, Hao, 089-2
Jia, Yue, 056, 068, 104
Jiang, Chunlian, 149
Jiang, Ning, 019
Jiang, Tao, 094
Jiang, Yang, 097
Jiang, Zhehan, 026, 104
Jiao, Hong, 102, 109, 116, 143, 150
Jin, Kuan Yu, 024
Jin, Yi, 038, 073
Jin, Ying, 057, 089-10
Johnson, Hayden, 101
Johnson, Janice Lee, 078
Johnson, Mark, 060, 153
Johnson, Matthew, 057, 139, 148
Jones, Andrew, 154
Jones, Paul Edward, 106
Jonson, Jessica L., 057
Joo, Sean, 077, 087, 109
Jorgensen, Terrence D, 166
Jozkowski, Kristen, 089-6
Julian, Marc W, 012, 032
Jung, Hyun Joo, 034
Jung, Ji Yoon, 103
Jung, Juyong, 110
Jurich, Daniel, 065-12

Kaliski, Pamela, 141
Kamata, Akhito, 011, 058, 085, 131
Kane, Michael, 107
Kang, Hyeon-Ah, 061, 077, 080, 118
Kannan, Priya, 005, 033, 079, 113
Kanopka, Klint, 054, 058, 089-14, 096, 144, 154
Kao, Shu-chuan, 168
Kaplan, David, 089-11
Kapoor, Radhika, 058, 096, 154
Kapoor, Shalini, 035
Kara, Yusuf, 131
Karamese, Hacer, 065-3, 065-14
Kares, Faith R, 112
Kartal, Gamze, 026
Katz, Daniel, 092
Kayton, Heather Leigh, 081, 087
Keefe, Janice, 089-5
Kehinde, Olasunkanmi, 081
Kelberlau, Darin, 164
Kell, Harrison, 075, 093
Keller, Lisa, 065-9, 091, 130
Keller-Margulis, Milena A., 065-2
Kendall Brooks, Lauren, 036
Keng, Leslie, 017, 039, 076, 120
Kennedy, Patrick, 052
Kern, Justin L., 081, 111, 154
Kerzabi, Emily, 020, 053
Ketterlin Geller, Leanne, 057, 071
Khorramdel, Lale, 103
Kilmen, Sevilary, 065-15, 089-5
Kilmen, Sevilay, 089-1
Kim, Ahyoung, 014
Kim, Brian H, 144
Kim, Dong-In, 032, 086
Kim, Hyung Jin, 065-6, 089-10, 154
Kim, Ji-Hye, 022
Kim, JongPil, 007, 013, 037
Kim, Kyung Yong, 073, 168
Kim, Minsung, 016
Kim, Seongeun, 145
Kim, Stella, 018, 097, 104, 121, 149
Kim, Sujin, 057
Kim, Sungyeun, 031
Kim, Sunhee, 089-16, 146, 154
Kim, Woomee, 057
Kim, Young Koung, 020, 058, 065-2, 154, 170
Kim, Youngwon, 089-12
Kim, Young Yee, 159
Kim, Yun-Kyung, 058
King, Sarah, 057
Kingsbury, G. Gage, 065-6
Kinsey, Devon, 124

Participant Index

(Last Name, First Name, Session Number)

Kintz, Tara, 162, 176
Klieger, David, 093
Knight, Melondy, 030
Ko, Amy, 097
Kobrin, Jennifer, 014, 105, 131
Koehn, Hans Friedrich, 018, 027
Kolen, Michael, 107
Kosh, Audra, 120
Koval, Jayma, 089-9
Kozikowski, Andrzej, 086
Kroehne, Ulf, 022
Kronberg, Carla, 057, 160
Kroopnick, Marc, 057, 089-10
Krost, Kevin, 089-5, 161
Kuhfeld, Megan, 069, 080
Kukea Shultz, Pohai, 030, 165
Kuklick, Livia, 072
Kumar, Lavanya Shravan, 087, 109
Kunina-Habenicht, Olga, 013, 065-7
Kuo, Tzu-Chun, 037, 065-13
Kuzinets, Vladimir, 085
Kwako, Alexander, 011
Kwon, Sunbeom, 018
Kyllonen, Patrick Charles, 053, 093

LaFlair, Geoff, 101, 130
Lai, Hollis, 063
Lai, Ka Wing, 018, 029
Lai, Viet, 085
Laitusis, Cara Cahalan, 156
Lakin, Joni, 012, 089-6
Lam, Jenny, 089-5
Lan, Andrew, 157
Landl, Erika, 050
Lane, John, 162, 176
Lane, Suzanne, 115
Lang, David, 144
Langi, Meredith, 065-6, 092, 169
Larson, Eric C., 011, 085
Laster, Christina, 164, 172
La Torre, Deborah, 081
Laukaityte, Inga, 149
Law, Evelyn, 080
Law, Nancy, 024
Lawless, Rene, 053
Lazarus, Sheryl, 065-1
LeBeau, Brandon, 037, 081
Lee, Boram, 089-10
Lee, Chansoon, 161
Lee, Daniel Yangsup, 020
Lee, Dayeon, 031
Lee, Dukjae, 091, 162
Lee, Eunji, 021, 031
Lee, Haeju, 073
Lee, Hollylynne, 147
Lee, Hyeryung, 010
Lee, Hyunjung, 019
Lee, Hyunsook, 073
Lee, Jade Caines, 055, 126, 160, 172, 175
Lee, Minho, 169
Lee, Sehee, 073
Lee, Seo Young, 031
Lee, Sooyong, 118
Lee, Sunhyoung, 021
Lee, Won-Chan, 026, 065-14, 065-6, 073, 089-10, 097, 104, 154
Lee, Yi-Hsuan, 056
Lee, Yongseok, 057
Lee, Yoonsun, 073
Lee Brown, Taneisha, 089-9
Leeming, Meghan, 073
Lehrfeld, Jon, 123, 137
Leino, Rosa, 057
Lembke, Erica, 057
Leng, Dihao, 103
Lentini, Jennifer, 020, 163
Leonard, Peter, 158
Leung-Gagné, Josh, 163
Leventhal, Brian C, 040, 062, 073, 080, 081, 123
Levine, Felice, 142
Levine Brown, Elizabeth, 089-11
Levy, Roy, 107
Lewis, Daniel, 046, 126, 136
Lewis, Jennifer L., 027, 140
Li, Chen, 152
Li, Guiyu, 143
Li, Isaac, 037
Li, Jie, 065-6
Li, Jie, 167
Li, Jiehan, 108
Li, Lanrong, 016, 028
Li, Linlin, 028
Li, Min, 054, 097
Li, Shuhong, 137
Li, Tongyun, 137
Li, Wen-Ching, 032
Li, Xin, 062, 091
Li, Xin (Grace), 065-3
Li, Xueming, 078
Li, Yalin, 016
Li, Zhen, 143
Li, Zhushan Mandy, 155
Li, Ziying, 057
Li, Zonggui, 065-10
Lian, Xu, 016
Liang, Min, 073
Liang, Qianru, 024, 061
Liang, Xinya, 089-11
Liao, Dandan, 102, 109
Liao, Manqian, 143
Liao, Xiangyi, 119
Liaw, Yuan-Ling, 054, 124, 166
Lim, Hwanggyu, 018, 077, 137
Lim, Jared, 097
Lim, Sangdon, 004, 169
Lim, Youn Seon, 027
Lin, Haiyan, 117
Lin, Jia, 131
Lin, Ye, 022
Lin, Zhongtian, 058
Lindner, Marlit Annalena, 072, 080
Lindner, Marlit, 072
Lindsay, Constance, 163
Ling, Guangming, 020, 057, 087, 123, 144, 155
Lissitz, Robert, 150
Liu, Allison, 118
Liu, Chunyan, 065-12, 109, 154
Liu, Guangyun, 154
Liu, Huan, 073
Liu, Jingchen, 061
Liu, Jinghua, 007, 148, 170
Liu, Juan, 016
Liu, Lei, 105, 124
Liu, Liu, 063
Liu, Lucia, 022
Liu, Lydia T., 130
Liu, Ou Lydia, 093, 123, 144
Liu, Ren, 026
Liu, Xiang, 065-19, 139
Liu, Xiangdong, 120
Liu, Xin Lucy, 089-17
Liu, Yang, 069
Liu, Ying, 132
Livne, Oren, 041, 139
Lo, Wen-Juo, 089-6
Locke, Victoria, 013, 170
Lockwood, J.R., 101, 149
Loken, Eric Eric, 061
Longe, Brendan, 090, 147, 169
Lopera-Oquendo, Carolina, 163
Lorie, William A., 167
Lottridge, Susan, 011, 039, 079, 085, 136 140, 157
Love, Quentin Ulysses Adrian, 017
Love, Quintin, 090
Lovelace, Temple S, 036
Lu, Ru, 025, 089-3
Lu, Yang, 117, 167
Lu, Yikai, 111
Lu, Ying, 057, 146
Lucas, Tracey Michelle, 057, 160
Luecht, Richard Melvin, 017, 065-16, 136, 146, 171
Lugu, Benjamin Kweku, 073
Luo, Jinwen, 147
Luo, Yachen, 128
Luo, Yong, 002, 009, 074, 123, 150
Lyons, Susan, 013, 055, 142
Lyu, Weicong, 111

Ma, Chenchen, 132
Ma, Cheryl, 065-16, 081
Ma, Jing, 065-14, 089-12
Ma, Wanjing Anya, 065-5
Ma, Wenchao, 027, 073, 132
Ma, Ye, 044, 083
MacGregor, David, 014
MacIntosh, Alexander, 065-17, 108
Madison, Matthew James, 001, 061, 087, 092, 138
Maeda, Hotaka, 097, 113
Maeda, Yukiko, 057
Maibach, Jacob, 080
Malatesta, Jaime, 169
Man, Kaiwen, 008, 012
Mangino, Tony A, 035
Manna, Venessa, 137
Mao, Xia, 085, 128, 168
Mardones, Constanza, 054
Marion, Scott, 014, 055, 110, 115, 163
Martin de los Santos Kleinz, Lisa, 147
Martineau, Joseph A., 150
Martinez, Alfonso, 065-7, 089-13, 096
Martinez, Jose Felipe, 054
Martínez-González, Adrián, 086
Martin-Raugh, Michelle, 020
Matta, Michael, 065-2
Matta, Tyler, 092
Mattison, Kate, 145

Participant Index
(Last Name, First Name, Session Number)

Mawhinney, Lynnette, 112
Maurs, Andreas, 147
Maxwell, Liam James, 057
Mbelede, Njideka Gertrude, 175
McCaffrey, Daniel, 057, 068, 140, 150
McCall, Martha, 019
Mcclure, Kyla, 076
McCormick, Carina M., 090
McCullough, Janeen, 070
McFadden, Mara, 155
McGrane, Joshua, 087
McHugh, Bridget, 028
McKlin, Tom, 089-9
McMahon, Margaret, 136
McMurtry, Teaira, 036
McNamara, Danielle, 152
McNeish, Daniel, 089-18
McVay, Aaron, 022
Medina Morales, Norma, 028
Mee, Janet, 011, 085, 113
Melaco, Carla, 112
Mendiola, Melchor Sanchez, 086
Meng, Huijuan, 044, 083
Mercer, Sterett H., 065-2
Merkle, Edgar, 054
Merriman, Jennifer, 064
Mertens, Ute, 072
Meyer, Jennifer, 073
Miao, Jing, 106
Michaelides, Michalis, 155
Michel, Rochelle, 156
Michels, Michael Andreas, 065-7, 113
Middlestead, Andrew J., 153, 164
Middleton, Kyndra, 030
Migunov, Igor, 023
Mikeska, Jamie, 085
Miller, Angela, 057
Miller, M. David, 089-6
Miller, Sherral, 050
Milligan, Sandra, 147
Mills, Christine, 067
Miranda, Alejandra, 089-4
Mislevy, Robert J., 107
Missall, Kristen, 145
Mitchell, Jamie, 131
Mo, Ya, 089-3
Mohammadi, Hamid, 011
Mojoyinola, Mubarak Olumide, 073, 081
Moncaleano, Sebastian, 057
Monroe, Scott, 077, 090
Montgomery, Melinda, 158, 167
Moore, Allen Christopher, 061
Moreno Luna, Arlyn Y, 089-9
Morrell, Monica, 069
Morrison, Kristin M., 065-4, 073, 126, 155
Moses, Tim, 020, 065-2, 146, 149, 170
Moskowitz, Joshua, 065-17
Moustaki, Irini, 089-2
Muckle, Timothy, 170
Mueller, Lorin, 097
Muenzen, Patricia, 141
Mugan, Leslie, 120
Mulvihill, Megan M, 105
Muniz, Jose, 171
Muniz, Jose, 171
Munson, Liberty, 106
Muntean, William J, 089-2
Murphy, Daniel, 090
Murphy, Victoria A, 081
Myers, Aaron, 090, 154
Myers, Matthew, 011, 065-2, 152

Nabors Olah, Leslie, 163
Nagel, Marie-Theres, 147
Nájera, Pablo, 027
Namsone, Dace, 097
Nascimento, Wallace, 015
Naveiras, Matthew David, 065-19, 089-12
Nemeth, Yvette M, 016
Nesbitt, Jaylin N, 079
Nese, Joseph F. T., 011, 085
Nguyen, Thien, 085
Niazi, Farhan, 065-7
Nickodem, Kyle, 147
Nie, Chang, 025
Nolan, Katherine, 109, 154
Nydick, Steven, 108, 149
Nye, Christopher, 087

Ober, Teresa, 065-4, 113
O'Donnell, Francis, 005, 065-9, 101
O'Dwyer, Eowyn, 105
Ogut, Burhan, 022
Oh, Hyeon-Joo, 007
Oh, Kyuseol, 021
Olgar, Suleyman, 084
Olivera Aguilar, Margarita, 163
Oliveri, Maria Elena, 050, 073, 101, 130, 147
Oluwalana, Olasumbo, 163
O'Neil, Harold, 088
O'Neill, Richard, 167
O'Neill, Thomas, 086
Ong, Thai Quang, 063
Onna, Marieke van, 149
O'Reilly, Tenaha, 152
Ormerod, Christopher, 085
O'Rourke, Hannah, 089-5
O'Rourke, Kevin P., 147
Ottmar, Erin, 022, 118
Ouyang, Jing, 132
Ouyang, Jinying, 026, 104
Ouyang, Wenli, 065-12
Ozkeskin, Emrah Emre, 131

Paccagnella, Marco, 051
Padgett, R Noah, 112
Padilla, Geraldo Bladimir, 073
Padro Collazo, Pascua, 089-9
Palermo, Corey, 011, 065-2
Palma, Jose R., 074
Pan, Qianqian, 024, 092
Pan, Tianshu, 167
Pan, Yiqin, 008, 143
Papanastasiou, Elena, 155
Pappas, Sandra, 028
Paris, Joseph, 144
Park, Elizabeth, 078
Park, Jung Yeon, 057, 160
Park, Seohee, 089-3
Parks, Charles, 088
Parra-Martinez, Fabio Andres, 057
Patarapichayatham, Chalie, 013
Patelis, Thanos, 064, 162
Patterson, Chris, 025
Patterson, Jim, 050
Patterson, Luke, 029
Patz, Richard, 108, 173
Peabody, Michael R, 029, 034, 112, 141
Pei, Bo, 113
Pellegrino, James, 082
Perez, Alexandra Lane, 057
Perez, Joselyn, 073
Perie, Marianne, 079
Pham, Duy N., 085
Pham, Paul, 097
Phenow, Aurore Yang, 086
Phenow, Aurore, 032
Phillip, kristy, 160
Phillip, Sharon, 057, 160
Phillippo, Kate, 089-11
Piacentini, Mario, 082
Pitoniak, Mary, 135
Plackner, Christie, 032
Pohl, Steffi, 051, 058
Por, Han-Hui, 082
Potgieter, Cornelis, 058, 094
Potter, Andrew, 011, 065-2
Powers, Sonya, 015, 037, 094
Prendez, Jordan Yee, 089-10
Price, Argenta, 082
Prihar, Ethan B, 118
Proctor, Thomas, 146
Puhan, Gautam, 119
Purpura, David, 145

Qazi, Nabeel, 170
Qi, Yi, 124
Qiao, Xin, 002, 009, 058, 065-19, 085, 131
Qin, Qi, 014
Qu, Yanxuan, 065-16
Quan, Jia, 011
Quan, Yale, 073
Quanbeck, Mari, 065-1
Quansah, Frank, 081
Quesen, Sarah, 007, 090
Quinones Perez, Isaris, 089-9
Quirk, Victoria L., 111

Rabe-Hesketh, Sophia, 022
Rabinowitz, Stanley N, 067, 105
Rafferty, Anna, 101
Rahal, Charles, 058
Ramsawak-Jodha, Nalini, 057, 160
Randall, Jennifer, 175
Rao, Analía, 089-9
Ravelo, Guillermo, 103
Raymond, Krystina, 097
Reckase, Mark, 107
Redman, Elizabeth, 042, 088
Reichert, Frank, 024
Reilly, Amy, 158
Ren, Hao, 117
Ren, Jinglei, 150
Rhemtulla, Mijke, 058
Rijmen, Frank, 099
Rikoon, Sam, 163
Rios, Joseph A., 030, 072, 073, 089-13, 096, 155, 165, 175

Participant Index

(Last Name, First Name, Session Number)

Robar, Julianne, 125
Robb, Colleen, 065-17
Roberts, Jeremy, 042, 088
Robin, Frederic, 056, 114
Rodriguez, Karina, 036
Rodriguez, Michael C., 028, 065-15, 075, 089-4, 138, 156, 174
Roebeck, Mark, 170
Roeber, Edward Dean, 176
Rollins, Taj, 161
Rome, Logan, 057, 065-17
Rosales de Véliz, Leslie Vanessa, 175
Rosca, Oxana, 081
Rosen, Yigal, 130
Rosenberg, Sharyn, 067
Ross, Linette P., 109, 154
Rotou, Ourania, 140
Roussos, Louis, 014, 054, 116
Rozunick, Chris, 158
Ruan, Chunyi, 152
Rubright, Jonathan, 145
Ruiz-Primo, Maria Araceli, 054
Runge, Andrew, 101
Runyon, Christopher, 011, 062
Russell, Michael, 089-4, 115, 138, 165
Rust, Keith, 068
Rutkowski, David, 103, 114, 166
Rutkowski, Leslie, 077, 114, 166
Ryu, Ehri, 065-10

Sabatini, John, 082
Saeidzadeh, Seyedehtanaz, 089-5
Sahin, Fusun, 057
Salas, Jorge, 089-12
Sales, Adam C, 022, 118
Sambucharan-Mohammed, Murella, 057, 160
Sammit, George, 085
Sanchez, Edgar, 075, 121
Sanders, Elizabeth A., 089-12
Sarac, Merve, 095, 109
Sarraf, Shimon, 147
Satkus, Paulius, 109, 145, 169
Sato, Edynn, 160
Sato, Yoshikazu, 023
Sayin, Ayfer, 029, 065-18, 131
Schaefer, Katarina, 073, 155
Schaller, Nils-Jonathan, 073
Schellman, Madeline, 065-8, 074, 147
Schmidt, Amy Elizabeth, 084
Schmidt, Amy, 084
Schneider, Christina, 060, 094, 136
Schneider, Wei S., 146
Schöber, Christian, 072
Schoemann, Alexander M, 166
Schonberg, Christina, 089-18, 145
Schor, David, 093
Schultz, Matthew, 029
Schumacker, Randall, 073
Schwartz, Robert, 072
Sgammato, Adrienne, 068, 129
Shafer Willner, Lynn, 112
Shapovalov, Yelisey A, 129, 162
Sharairi, Sid, 013
Shaw, Emily, 050
Shear, Benjamin R., 061
Shen, Yawei, 099, 167
Sheng, Yanyan, 065-13
Shepard, Lorrie Ann, 067
Shermis, Mark David, 029, 085
Shetty, Sandeep, 029
Shi, Dexin, 026
Shi, Qingzhou, 051, 089-6
Shibayama, Tadashi, 023
Shih, Ching-Lin, 025
Shin, David, 081, 099, 117, 137, 167
Shin, Hyo Jeong, 114
Shin, Jinnie, 089-1
Shin, Minkyoung, 073
Shin, Nami, 097
Shivraj, Pooja, 170
Shono, Yusuke, 089-14
Shrestha, Shovana, 089-5
Sikali, Emmanuel, 006, 123, 166
Sinharay, Sandip, 033, 039, 065-16, 109, 119, 148
Sinval, Jorge, 054
Sireci, Stephen G, 022, 052, 057, 064, 091, 101, 120, 130, 150, 164, 171
Sitarenios, Gill, 065-17
Skoblow, Hanamori, 100
Skorupski, William, 065-8, 143
Slater, Sharon, 140
Smith, Larissa, 168
Smith, Thomas M., 110
Snow, Erica, 124
Soland, James, 069, 096
Solano-Flores, Guillermo, 030, 071, 156
Someshwar, Shonai, 021
Somsong, Sarunya, 085
Song, Jieqing, 092
Song, Yi, 105, 152
Song, Yoon Ah, 065-3
Song, Yuting, 089-5
Sonnleitner, Philipp, 065-7, 113
Sorrel, Miguel A., 027
Sotelo, Jose de Jesus, 057
Springer, Matthew, 110
Stark, Stephen, 087, 109
Steedle, Jeffrey, 030, 035, 060
Steimel, Kenneth, 124
Steinberg, Jonathan, 085
Stewart, John, 016
Stoeger, Jordan Nelson, 065-8, 129
Stopek, Joshua, 029
Stout, William, 122
Struthers, Vince, 086
Student, Sanford, 019, 076, 090
Su, Shiyang, 052
Su, Yu-Lan, 020, 106
Suárez-Álvarez, Javier, 091, 101
Subedi, Dipendra, 016, 099
Subhiyah, Raja G, 065-12
Suh, Hongwook, 073
Suh, Yon Soo, 100
Suh, Yon Soo, 092
Suk, Youmi, 087
Sum, Jane, 080
Sun, Sundance Zhihong, 147
Sussman, Joshua, 116, 128
Svensson, Cicek, 064
Svetina Valdivia, Dubravka, 077, 089-4, 147
Swain, Matthew, 089-10
Swaminathan, Hariharan, 171
Sweet, Tracy, 065-19
Swygert, Kimberly, 024, 065-12

Talreja, Vinita, 044, 083
Tan, Bin, 018, 022, 051, 113
Tan, Sinon, 143
Tan, Xuan (Adele), 065-15, 093, 147, 170
Tan, Yanyan, 167
Tang, Nai-En, 065-11
Tang, Steven, 143
Tang, Xiuxiu, 094
Tao, Shuqin, 013, 028, 034
Teitelbaum, Jeremy, 061
Telfort, Roseline, 081
Templin, Jonathan, 065-7, 094
Thiessen, Brad, 129
Thomas, Elizabeth R., 057
Thomas, Jay, 089-8, 108
Thompson, Kathryn Nicole, 060, 081
Thompson, William Jacob, 014, 087, 116
Thompson, W. Jake, 020, 092
Thum, Yeow Meng, 090, 128, 150
Thurlow, Martha, 065-1, 089-7
Tian, Chen, 102
Tijmstra, Jesper, 114
Tingir, Seyfullah, 016, 065-19
Todd, Jessica Andrews, 015
Tolentino, Lissette, 018
Tomkowicz, Joanna, 032, 086
Tong, Ye, 066, 131, 142
Topczewski, Anna, 026, 070
Topczewski, Anna, 022
Toptas, Cigdem, 159
Torres, Chloe, 153
Torres Irribarra, David, 138
Toton, Sarah Linnea, 008, 148
Toyama, Yukie, 173
Traynor, Anne, 155
Tsai, Chia-Lin, 065-11
Tu, Naidan, 087, 109
Tubbs Dolan, Carly, 147
Tufail, Mubeshera, 061, 147
Turner, Ronna, 035, 089-6
Twing, Jon S, 084
Tyack, Lillian, 103

Ulitzsch, Esther, 048, 051, 096
Underwood, Sarah, 084

Valdivia Medinaceli, Montserrat B, 077, 095
Vanacore, Kirk P, 022, 118
Van der Linden, Wim J, 171
Van Rijn, Peter, 056, 119, 169
Vasquez-Colina, Maria, 086
Veldkamp, Bernard, 048
Verges, Vince, 153
Vernon, Annette, 065-9
Vispoel, Walter, 010, 028
Vo, Thao, 142
Vo, Yen, 065-9
von Davier, Alina A, 039, 053, 130, 149, 168
von Davier, Matthias, 056, 103
Vorenkamp, Ellen, 176

Participant Index

(Last Name, First Name, Session Number)

Walker, Marcus, 063
Walker, Michael E., 039, 106
Walkowiak, Templ A., 129
Wall, Nathan, 045
Wallace, Leslie, 068
Wallin, Gabriel, 035, 089-2
Walsh, Cole, 108
Walters, Kristy, 176
Wan, Ping, 086
Wan, Siyu, 065-9, 154
Wang, Aijun, 097
Wang, Bowen, 111
Wang, Changjiang, 016, 099
Wang, Chuang, 149
Wang, Chun, 052, 073, 089-16, 132, 151
Wang, Chunxin, 065-6
Wang, Dongran, 012
Wang, Dongwei, 065-13, 091
Wang, Huan, 032
Wang, Jincai, 149
Wang, Kuo, 085
Wang, Nixi, 054, 080
Wang, Selena, 102
Wang, Serena, 130
Wang, Shichao, 035, 154
Wang, Shiyu, 099, 117, 132
Wang, Shudong, 120, 143
Wang, Songtao, 097
Wang, Ting, 086
Wang, Xi, 065-14, 089-12
Wang, Xiaolin, 106
Wang, Yihao, 011
Wang, Yu, 027, 116
Wang, Yuan, 020, 053, 078
Wang, Yurou, 028
Wang, Zhuoran, 089-2
Warner, Zachary, 153, 164
Way, Denny, 146, 154
Webb, Blue, 006
Weber, Anjali, 083
Weeks, Jonathan, 057, 097, 119, 159
Weese, James, 035
Wei, Xin, 051, 080
Wei, Youhua, 119
Weir, John, 063
Weiss, David, 052
Weissman, Alexander, 058, 070
Welch, Catherine, 065-14, 065-9, 073, 081, 089-12, 089-15
Wellberg, Sarah, 076
Wells, Craig, 065-13
Welsh, Megan, 110
Wheeler, Jordan M., 065-8
Wheeler, Kelley, 086
Whitmer, John, 151, 157
Whittaker, Tiffany, 118
Wiberg, Marie, 149
Wibowo, Arianto, 157
Wieman, Carl, 082
Wiley, Andrew, 064
Wilhelm, Anne G., 129
Williams, Kevin, 057, 123
Wilson, Jonee, 129
Wilson, Joshua, 011, 065-2
Wilson, Mark, 023, 072, 173
Wind, Stefanie A., 028, 089-6, 138
Wine, Marjorie, 060
Winter, Sonja D, 100
Winter, Sonja D, 054
Winters, Erin, 089-4
Wise, Steven, 065-6, 072, 080, 096
Witmer, Sara, 089-5
Wolf, Mikyung Kim, 089-7
Wollack, James, 109, 148
Wong, Yun Leng, 052
Wongvorachan, Tarid, 018, 022, 029, 096, 113
Woo, Ada, 067
Woo, Yejin, 031
Woods, Carol, 094
Workman, Trent, 158, 164
Wraga, Bill, 091
Wu, Amery, 018
Wu, Yi-Chen, 065-1, 089-7
Wu, Yi-Fang, 045, 058, 112
Wyse, Adam E, 017, 086

Xi, Nuo, 068
Xiao, Houping, 099
Xiao, Jiaying, 089-16, 132, 169
Xiao, Xingyao, 022
Xie, Aolin, 070
Xie, Benjamin, 097
Xin, Tao, 015, 025
Xiong, Jiawei, 102
Xiong, Yao, 089-18, 145, 151
Xu, Gongjun, 089-16, 132
Xu, Guanlan, 108
Xu, Lingling, 026, 104
Xu, Menglin, 019
Xu, Rujun, 096
Xu, Xiangyu, 113
Xu, Xiaochen, 145

Yablonski, Maya, 131
Yan, Duanli, 043, 066, 152
Yan, Wei Wei, 065-17
Yan, Yan, 054
Yancey, Kevin, 101
Yang, Hongyu, 081
Yang, Ji Seung, 069
Yang, Zhitong, 020
Yao, Lihua, 038, 063
Yavuz, Sinan, 006, 166
Yavuz Temel, Güler, 025
Ye, Daisy, 012
Yeatman, Jason, 065-5, 089-14, 131
Yen, Shu Jing, 089-15
Yeum, Sung Kun, 031
Yigit, Hulya Duygu, 061
Yildirim-Erbasli, Seyma N., 051, 063, 124, 140
Ying, Zhiliang, 061
Yoo, Hanwook, 007, 089-7, 137, 170
Young, Mackenzie, 079
Younger, Jessica, 088
Yuan, Ye, 065-8
Yumsek-Akbaba, Meltem, 024

Zahner, Doris, 156
Zan, Yixin, 057
Zapata-Rivera, Diego, 005, 033, 043
Zavitkovsky, Paul, 128
Zeng, Lin, 081
Zenisky, April 005, 033, 101, 171
Zhang, Fan, 065-2
Zhang, Jihong, 026, 065-7, 080
Zhang, Jinming, 081, 089-17
Zhang, Lijin, 089-11
Zhang, Litong, 012
Zhang, Mengyao, 168
Zhang, Mingqin, 081, 089-15
Zhang, Mo, 066, 097
Zhang, Oliver, 146
Zhang, Susu, 018, 051, 080, 108
Zhang, Ting, 006, 123, 166
Zhang, Tongxin, 015
Zhang, Wenqing, 016
Zhang, Xiao, 012
Zhang, Yichi, 159
Zhang, Yifan, 038, 073
Zhang, Yu, 097
Zhang, Zhonghua, 023, 147
Zhao, Mingye, 085
Zhao, Xinchu, 124
Zhao, Yang, 089-3, 145
Zhao, Yu (Tracy), 117
Zheng, Bin, 012
Zheng, Chanjin, 016
Zheng, Guoguo, 016, 065-19
Zheng, Xiaying, 159
Zheng, Yi, 108
Zhong, Xiaoting, 037
Zhou, Junfan, 065-4
Zhou, Todd, 143
Zhou, Xuechun, 089-17
Zhu, Danqi, 034, 077
Zhu, Ruoyi, 057
Ziedzor, Reginald, 016
Ziedzor, Reginald, 065-19
Ziegler, Garrett, 050
Zilberberg, Anna, 015
Zor, Selay, 061, 087
Zou, Zhimin, 038
Zu, Jiyun, 078
Zumbo, Bruno D., 171

Participant Emails

(Last Name, First Name; Affiliation; Email)

.,Ketan; University of Massachusetts; ketan@umass.edu
Aanjaneya, Mridul; Rutgers University; ma635@rutgers.edu
Abad, Francisco J.; Universidad Autonoma de Madrid; fjose.abad@uam.es
Abdalla, Widad; widad.abdalla1@upr.edu
Abebe, Rediet; UC, Berkeley; abebe@berkeley.edu
Abulela, Mohammed; mhady001@umn.edu
Abuosbeh, Zein; University of Toronto; zein.abuosbeh@mail.utoronto.ca
Ackerman, Terry; University of Iowa; terryackerman2@gmail.com
Adams, Elizabeth L.; Southern Methodist University; beth.greive@gmail.com
Ahadi, Stephan; Cambium Assessment, Inc; stephan.ahadi@gmail.com
Ahn, Dahwi; Iowa State University; dahn@iastate.edu
Ahn, Jungwon Rachael; ; jahn20@fordham.edu
Ahn, Sunyoung; Jeonghwa Arts College; mistral21@hanmail.net
Akande, Christiana Aikenosi; National Commission On Certification of Physician Assistants; christianaa@nccpa.net
Alahmadi, Sarah; James Madison University; alahmasi@jmu.edu
Alakbarova, Vafa; ; valakbarova@umass.edu
Albano, Tony; University of California, Davi; adalbano@ucdavis.edu
Ali, Usama; Educational Testing Service; uali@ets.org
Alkhalaiwi, Alaa; Morgan State University; alalk4@morgan.edu
Allen, Jeff M.; ACT, Inc; Jeff.Allen@act.org
Almond, Russell G; Florida State Univeristy; ralmond@fsu.edu
Aloe, Ariel M.; University of Iowa; ariel-aloe@uiowa.edu
Alonzo, Alicia; Michigan State University; alonzo@msu.edu
Alvero, A J; Stanford University; ajalvero@stanford.edu
Amissah, Christopher Martin; Morgan State University; chami2@morgan.edu
An, Lily; Harvard Graduate School of Education; lily_an@g.harvard.edu
An, Xiaozhu; IXL Learning; xan@ixl.com
Andrade, Heidi L.; University at Albany; handrade@albany.edu
Andrade-Lotero, Luis Alejandro; Pearson; alejandro.andradelotero@pearson.com
Andress, Tim; The University of Texas at Austin; tim.andress@utexas.edu
Andrews, Benjamin; Inteleos; benjamin.andrews@inteleos.org
Anghel, Ella; ; anghel@bc.edu
Anguiano-Carrasco, Cristina; ; cristina.anguiano-carrasco@act.org
Ankomah, Francis; University of Cape Coast; francis.ankomah@stu.ucc.edu.gh
Antal, Judit; College Board; jantal@collegeboard.org
Anyidoho, Abena; The Ohio State University; anyidoho.1@osu.edu
Araneda, Sergio; University of Massachusetts Amherst; saraneda@umass.edu
Arce Nazario, Rafael; University of Puerto Rico- Rio Piedras; rafael.arce@upr.edu
Arcilla, Mary Anne; Educational Testing Service; marcilla@ets.org
Armstrong, Maria; Association of Latino Administrators and Superintendents; maria@alasedu.org
Arslan, Burcu; Educational Testing Service Global B.V.; barslan@ets.org
Arthur, Ann; ACT; ann.arthur@act.org
Asamoah, Nana Amma; University of Arkansas; nbasamoa@uark.edu
Avvisati, Francesco; OECD; francesco.avvisati@gmail.com
Awwal, Nafisa; University of Melbourne; n.awwal@unimelb.edu.au
Ayers-Wright, Elizabeth; Cambium Assessment; eaa18sb@gmail.com
Ayik, Bilgehan; George Mason University; bayik@gmu.edu
Babcock, Ben; Association for Materials Protection and Performance; babco062@umn.edu
Badrinarayan, Aneesha; Learning Policy Institute; abadrinarayan@learningpolicyinstitute.org
Baghestani, Shireen; WIDA Consortium; shireenb@iastate.edu
Bahr, Jan Luca; ; bahr@leibniz-ipn.de
Baidoo-Anu, David; ; baidooanu.david@queensu.ca
Baig, Basim; Duolingo; basim@duolingo.com
Bailey, Paul; American Institutes for Research; pbailey@air.org
Baker, Doris Luft; The University of Texas at Austin; doris.baker@austin.utexas.edu
Baker, Eva L.; UCLA; eva@ucla.edu
Bakken, Sara; BrainPop; sarab@brainpop.com
Baldwin, Peter; National Board of Medical Examiners; pbaldwin@nbme.org
Ballard, Laura D; Educational Testing Service; lballard@ets.org
Bandalos, Deborah; James Madison University;
Banerjee, Sube; University of Plymouth; sube.banerjee@plymouth.ac.uk
Bao, Yu; James Madison University; bao2yx@jmu.edu
Barragan Torres, Mariana; IWERC, University of Illinois; marianab@uillinois.edu
Barry, Carol L; American Board of Surgery; cbarry@absurgery.org
Bartz, Kayla; Michigan State University; bartzkay@msu.edu
Basaraba, Deni; Amplify Education; dbasaraba@amplify.com
Bashkov, Bozhidar M.; IXL Learning; bbashkov@ixl.com

Participant Emails

(Last Name, First Name; Affiliation; Email)

Bates, Meg; IWERC, University of Illinois; megbates@illinois.edu
Bay, Luz; College Board; lbay@collegeboard.org
Becker, Kirk; Pearson; kirk.becker@pearson.com
Becker, Rachael N; Southern Methodist University; rnbecker@mail.smu.edu
Bediwy, Ahmed; The University of Iowa; ahmed-bediwy@uiowa.edu
Bei, Ni; University of Washington; nbei@uw.edu
Beilstein, Shereen Oca; IWERC, University of Illinois; beilste2@uillinois.edu
Beimers, Jennifer; Pearson; jennifer.beimers@pearson.com
Beiting-Parrish, Magdalen; CUNY Graduate Center; mbeiting@gradcenter.cuny.edu
Bell, Nathan; American Educational Research Association; nbell@aera.net
Belov, Dmitry I.; Law School Admission Council; dbelov@lsac.org
Belwalkar, Bharati; AIR; bbelwalkar@air.org
Bennett, Randy; ETS; rbennett@ets.org
Berenbon, Rebecca; ; berenbon.1@osu.edu
Bergner, Yoav; ; yoav.bergner@nyu.edu
Berry, Yufeng; Minnesota Department of Education; yufeng.berry@state.mn.us
Beswick, Bruce; The University of Melbourne; bruceab@unimelb.edu.au
Betts, Joe; NCSBN; jbetts@ncsbn.org
Bezirhan, Ummugul; Boston College TIMSS & PIRLS International Study Center; bezirhan@bc.edu
Biancarosa, Gina; University of Oregon; Ginab@uoregon.edu
Binici, Salih; ; salih.binici@fldoe.org
Binzak, John Vito; Pearson Clinical Assessment; john.binzak@pearson.com
Bishop, Kyoungwon Lee; WIDA at University of Wisconsin Madison; kei.bishop2@gmail.com
Blackman, Horatio; National Urban League; hblackman@nul.org
Boatman, Angela; Boston College; boatmana@bc.edu
Bolsinova, Maria; Utrecht University; m.a.bolsinova@uu.nl
Bolt, Daniel; University of Wisconsin - Madison; dmbolt@wisc.edu
Bonifay, Wes; University of Missouri; bonifayw@missouri.edu
Borgonovi, Francesca; University College London; Organisation for Economic Co-Operation and Development; f.borgonovi@ucl.ac.uk
Borowiec, Katrina; Boston College; katrina.borowiec@bc.edu
Bowe, Anica; Oakland University; bowe@oakland.edu
Boyd, Aimee; Curriculum Associates; aimeeboyd@cainc.com
Boyd, Brittany; American Institutes for Research; bboyd@air.org
Bradford, Lydia; Michigan State University; bradf134@msu.edu
Bradshaw, Laine; Pearson; lainebradshaw@gmail.com
Bratsch-Hines, Mary; University of Florida; bratsch@coe.ufl.edu
Brennan, Robert; University of Iowa; robert-brennan@uiowa.edu
Brice, Amanda; Curriculum Associates; abrice@cainc.com
Bridgeman, Brent; ETS; bbridgeman@ets.org
Briggs, Derek Christian; University of Colorado Boulder; derek.briggs@colorado.edu
Britton, Tolani; UC, Berkeley; tabritton@berkeley.edu
Brooks, Christopher; UNC; chbrooks@live.unc.edu
Brown, Alaysia Marie; Educational Testing Service (ETS); ambrown@fas.harvard.edu
Brown, Gavin T. L.; The University of Auckland; gt.brown@auckland.ac.nz
Brown, Naomi; George Mason University; nbrown27@gmu.edu
Brown, Richard; National Math and Science Initiative; rbrown@nms.org
Brunetti, Matthew; WestEd; matthew.brunetti@gmail.com
Bryer, Jason; City University of New York; jason@bryer.org
Buchholz, Janine; DIPF | Leibniz Institute for R; buchholz@dipf.de
Buckendahl, Chad W.; ACS Ventures, LLC; cbuckendahl@acsventures.com
Buckley, Jack; Roblox Corp; jbuckley@roblox.com
Bulut, Cigdem; ; hcyavuz@gmail.com
Bulut, Okan; University of Alberta; bulut@ualberta.ca
Burchell, Diana; University of Toronto; diana.burchell@mail.utoronto.ca
Burgess, Yin; National Registry of EMTs; yburgess@nremt.org
Burgmanis, Girts; ; girts.burgmanis@lu.lv
Burke, Matthew; AWS; mjburke@amazon.com
Burkhardt, Amy; Stanford University; burkhardt.amy@gmail.com
Burstein, Jill; Duolingo; jill@duolingo.com
Butterbaugh, Donna J; (ISC)2; butte009@umn.edu
Büyükkıdık, Serap; Ohio State University; sbuyukkidik@gmail.com
Buzo-Casanova, Enrique; Universidad Nacional Autónoma de México (UNAM); enrique_buzo@cuaieed.unam.mx
Bynum, Bethany; HumRRO; bbynum@humrro.org
Cai, Li; UCLA; lcai@ucla.edu
Cai, Liuhan Sophie; Cognia; cliuhan@gmail.com
Caliço, Tiago A.; Independent; tcalico@protonmail.com

Participant Emails

(Last Name, First Name; Affiliation; Email)

Calliouet, Ruth; Louisiana Department of Education; Ruth.Caillouet@la.gov
Camara, Wayne J.; LSAC; waynecamara@gmail.com
Camargo Salamanca, Sandra Liliana; Purdue University; camargos@purdue.edu
Canbolat, Yusuf; Indiana University Bloomington; ycanbola@iu.edu
Cancado, Luciana; Curriculum Associates; lcancado@cainc.com
Canto, Phil; phil.canto@fldoe.org
Cao, Canxi; Beijing Normal University; ccx0918@outlook.com
Cao, Yi; Educational Testing Service; ycao@ets.org
Cappaert, Kevin; Curriculum Associates; kcappaert@cainc.com
Cardoso-Leite, Pedro; University of Luxemburg; pedro.cardosoleite@uni.lu
Cardoza, Daniela; The University of Iowa; daniela-cardoza@uiowa.edu
Cardwell, Ramsey; Duolingo; ramsey@duolingo.com
Carr, Peggy; National Center for Education Statistics; peggy.carr@ed.giv
Carroll Miranda, Joseph; University of Puerto Rico- Rio Piedras; joseph.carroll@upr.edu
Carroll, James; ETS; jcarroll@ets.org
Cartier, Salenah; ETS; scartier@ets.org
Casabianca-Marshall, Jodi; Educational Testing Service; jcasabianca@ets.org
Casas, Maritza; University of Massachusetts Amherst; mcasas@umass.edu
Castaneda, Ruben; College Board; rcastaneda2@ucmerced.edu
Chalmers, R. Phillip; York University; chalmrp@yorku.ca
Cham, Heining; Fordham University; hcham@asu.edu
Cham, Heining; Fordham University; hcham@fordham.edu
Chan, Jason C.K.; Iowa State University; ckchan@iastate.edu
Chang, Ammi; Purdue University; chang734@purdue.edu
Chang, Hua-Hua; Purdue University; chang606@purdue.edu
Chang, Kuo-Feng; ; kuo-feng-chang@uiowa.edu
Chang, Wanchen; Cambium Assessment; claire.chang@cambiumassessment.com
Chao, Hsiu-Yi; National Taiwan Ocean University; hsiuyi1118@gmail.com
Chavez, Carlos; University of Minnesota - Twin Cities; chave143@umn.edu
Chawla, Kamal; The Collegeboard; kamalc@udel.edu
Chen, Becky; University of Toronto; xi.chen.bumgardner@utoronto.ca
Chen, Cheng Te; National Tsing-Hua University, Taiwan; chengte@mx.nthu.edu.tw
Chen, Dandan (Danielle); University of Illinois, Urbana-Champaign; dandan.c.chen@gmail.com
Chen, Hong; University of Iowa; hchen102@uiowa.edu
Chen, Jianshen; College Board; cachen@collegeboard.org
Chen, Jie; Measurement Incorporated; jcmuku@gmail.com
Chen, Jihang; Boston College; jihang@bc.edu
Chen, Jing; Cambium Assessment; jingchen2022@gmail.com
Chen, Jinsong; The University of Hong Kong; jinsong.chen@live.com
Chen, Jyun-Hong; National Cheng Kung University; psyjhc@gs.ncku.edu.tw
Chen, Ofer; New York University; oc587@nyu.edu
Chen, Troy; Measurement Incorporated; TChen@measinc.com
Chen, Xiaowen; George Mason University; xchen29@gmu.edu
Chen, Xing; Fordham University; xchen358@fordham.edu
Chen, Yinghan; University of Nevada, Reno; yinghanc@unr.edu
Chen, Yunxiao; London School of Economics and Political Science; y.chen186@lse.ac.uk
Cheng, Yiling; Kaohsiung Medical University; yilingcheng@kmu.edu.tw
Cheng, Ying; University of Notre Dame; ycheng4@nd.edu
Chiang, Yi-Chen; NABP; ychiang@nabp.pharmacy
Chien, Yuehmei; ; yuehmeir@gmail.com
Chiu, Chia-Yi; University of Minnesota; cchiu@umn.edu
Chiu, Ming Ming; Education University of Hong Kong; mingmingchiu@gmail.com
Cho, Sun-Joo; Peabody College of Vanderbilt; sj.cho@vanderbilt.edu
Cho-Baker, sugene; ETS; SCHOBAKER@ets.org
Choe, Edison M.; Renaissance; edison.choe@renaissance.com
Choi, Hunwon; Ewha Womans University; hwon0208@ewhain.net
Choi, Hye-Jeong; Human Resources Research Organization; hchoi@humrro.org
Choi, Ikkyu; ; ichoi001@ets.org
Choi, Jaehwa; George Washington University; jaechoi@gwu.edu
Choi, Jinah; Edmentum, Inc.; Jinah.Choi@edmentum.com
Choi, Jinnie; Savvas Learning Company; jinnie.choi@gmail.com
Choi, Kilchan; CRESST/UCLA; kcchoi@ucla.edu
Choi, Seung W.; University of Texas at Austin; schoi@austin.utexas.edu
Choi, Youn-Jeng; Ewha Womans University; younjengchoi@ewha.ac.kr
Christie, Thomas; NWEA; thomas.christie@nwea.org
Christopherson, Sara; Wisconsin Center for Education Products & Services- UW Madison; sara.christopherson@wceps.org
Chuah, Siang Chee; College Board; dchuah@collegeboard.org

Participant Emails

(Last Name, First Name; Affiliation; Email)

Chung, Greg; ; greg@ucla.edu
Chung, Gregory; ; chung@cresst.org
Chung, Jinmin; University of Iowa; jinmin-chung@uiowa.edu
Cirlot, La'Shea; Cognia; la'shea.cirlot@cognia.org
Cisterna-Alburquerque, Dante; ETS; dcisterna@ets.org
Cizek, Gregory; University of North Carolina at Chapel Hill; cizek@unc.edu
Clark, Amy; ATLAS: University of Kansas; akclark@ku.edu
Clark, Shannon; University of Georgia; mallory.clark@uga.edu
Clauser, Amanda; National Board of Medical Examiners; aclauser@nbme.org
Clauser, Brian; National Board of Medical Examiners; bclauser@nbme.org
Clauser, Jerome; American Board of Internal Med; jclauser@abim.org
Clementz, Annie Rae; Illinois State Board of Education; aclement@isbe.net
Close, Catherine; Renaissance Learning; catherine.close@renaissance.com
Close, Catherine; Renaissance Learning; cathynclose@gmail.com
Cobb, Paul; Vanderbilt University; paul.cobb@vanderbilt.edu
Coggeshall, Whitney Smiley; Educational Testing Service (ETS); whitknee48@gmail.com
Cohen, Allan; University of Georgia; acohen@uga.edu
Coles, Jessica; Measurement Incorporated; jcoles@measinc.com
Collard, Jasmine; University of Notre Dame; jcollard@nd.edu
Colvin, Kimberly; University at Albany, SUNY; kcolvin@albany.edu
Cook, Carson; NWEA; carson.cook@nwea.org
Cook, Robert; Cognia; robert.cook@cognia.org
Cook, Robert J.; Cognia, Inc.; rob.cook.online@gmail.com
Corazza, Luke; IXL Learning; lcorazza@ixl.com
Çorbacı, Ergün Cihat; Gazi University; e.cihat.corbaci@gmail.com
Corrado, Kelly; PBS KIDS; kncorrado@pbs.org
Corrin, Linda; Deakin University; linda.corrin@deakin.edu.au
Cox, Olivia; University of Colorado Boulder; olivia.cox@colorado.edu
Crawford, Brandon; Indiana University; brancraw@iu.edu
Crespo Cruz, Eduardo Javier; UMass Amherst; ecrespocruz@umass.edu
Crinion, Miriam; Buros Center for Testing - University of Nebraska-Lincoln; miriam.crinion@huskers.unl.edu
Croft, Michelle; ; croft.michelle@gmail.com
Cruce, Ty; ACT; Ty.cruce@act.org
Cruz, Tania; University of Delaware; taniacc@udel.edu
Cuhadar, Ismail; Ministry of Education, Turkey; ismail.cuhadar@meb.gov.tr
Cui, Mengyao; Cambium Assessment, Inc.; mengyao.cui@cambiumassessment.com
Cui, Weiwei; College Board; wcui@collegeboard.org
Cui, Wenju; ETS; wcui@ets.org
Cui, Ying; University of Alberta; yc@ualberta.ca
Culpepper, Steven; University of Illinois at Urba; sculpepp@illinois.edu
Curnow, Christina; AIR; ccurnow@air.org
Custer, Michael; Riverside Insights; michael.custer@riversideinsights.com
Dadey, Nathan; Center for Assessment; ndadey@nciea.org
Dai, Shenghai; Washington State University; s.dai@wsu.edu
Daisher, Ted; Curriculum Associates; TDaisher@cainc.com
Dallas, Andrew; National Commission On Certification of Physician Assistants; drewd@nccpa.net
Damico, Danielle; Amplify Education; ddamico@amplify.com
Darling-Hammong, Linda; Learning Policy Institute; ldh@learningpolicyinstitute.org
Davis, Jennifer; Amazon Web Services (AWS); jedavisr@amazon.com
Davis, Laurie; Curriculum Associates; laurie@davistx.com
Davis-Becker, Susan; ACS Ventures, LLC; sdavisbecker@acsventures.com
Davison, Mark; University of Minnesota; mld@umn.edu
De Boeck, Paul; OSU; deboeck.2@osu.edu
de la Torre, Jimmy; University of Hong Kong; j.delatorre@hku.hk
De Lisle, Jerome; University of The West Indies,; jeromedelisle@yahoo.com
Deane, Paul; ETS; pdeane@ets.org
DeMars, Christine; James Madison University; demarsce@jmu.edu
Demir, Cihan; Washington State University; cihan.demir@wsu.edu
Demirkaya, Onur; Riverside Insights; onurdmrkaya@gmail.com
Denbleyker, Johnny; Kaplan; jdenblge@yahoo.com
Deng, Jiayi; University of Minnesota; deng0194@umn.edu
Deng, Jiayi; University of Minnesota, Twin Cities; jiayideng0726@gmail.com
Deng, Nina; Kaplan Inc.; nndeng@gmail.com
Denissov, Serguei; ETS; SDENISSOV@ets.org
Denner, Madelynn; University of Notre Dame; mdenner@nd.edu
Deribo, Tobias; DIFP | Leibniz Institute for Research and Information In Education; deribo@dipf.de
Devasia, Nimmi; Educational Testing Service; ndevasia@ets.org

Participant Emails

(Last Name, First Name; Affiliation; Email)

Devkota, Rashmi; University of Alberta; rdevkota@ualberta.ca

DeWeese, Joseph; University of Minnesota-Twin Cities; gree2903@umn.edu

Dhakal, Khagendra Raj; King Mongkut'S University of Technology Thonburi; raj.kmutt@gmail.com

Diao, Hongyu; Educational Testing Service (ETS); hdiao@ets.org

Diao, Qi; ETS; qdiao@ets.org

Dilek, Ismail; ; ismail-dilek@uiowa.edu

DiStefano, Christine; University of South Carolina; distefan@mailbox.sc.edu

Do, Tai; University of Minnesota Depart; Doxxx078@umn.edu

Doğuyurt, Mehmet Fatih; Gazi University; doguyurtfatih@gmail.com

Domingue, Ben; Stanford University; bdomingu@standford.edu

Domingue, Ben; Stanford University; ben.domingue@gmail.com

Dong, Dongsheng; University of Washington; dongsd@uw.edu

Dong, Yixiao; University of Denver; yixiao.dong@du.edu

Donoghue, John; Educational Testing Service; jrdonoghue@comcast.net

Dorsey, David; HumRRO; ddorsey@humrro.org

Doupe, Malcolm B.; University of Manitoba; MalcolmBray.Doupe@umanitoba.ca

Drame, Elizabeth; University of Wisconsin-Milwaukee; Erdrame@uwm.edu

Draney, Karen; University of California, Berk; kdraney@berkeley.edu

Dresher, Amy; ETS; adresher@ets.org

Duan, Qizhou; University of Notre Dame; qduan@nd.edu

Duan, Yinfei; University of Alberta; yinfei1@ualberta.ca

Dumas, Denis; University of Georgia; Denis.Dumas@uga.edu

Dunbar, Stephen; Universiy of Iowa; steve-dunbar@uiowa.edu

Dunn, Jennifer; Pearson; jenn.dunn@pearson.com

Durrence, Debbie; Gwinnett County Public Schools; debbie.durrence@gcpsk12.org

Dwyer, Andrew; American Board of Pediatrics; adwyer@abpeds.org

Dwyer, Kevin; Cambium Assessment; kevin.dwyer@cambiumassessment.com

Dymchuk, Emily; University of Alberta; dymchuk@ualberta.ca

Eacker, Halley; Measurement Incorporated; heacker@measinc.com

Ece, Berivan; Northwestern University; berivan.eceusta@northwestern.edu

Eckerly, Carol; Educational Testing Service; ceckerly@ets.org

Edi, Daniel; Pearson Assessments and Qualifications; daniel.edi@pearson.com

Edwards, Douglas; Georgia Institute of Technology; doug.edwards@ceismc.gatech.edu

Edwards, Kelly; University of Virginia; kde2cp@virginia.edu

Egan, Karla; Edmetric LLC; karla.egan@edmetric.com

El Masri, Yasmine; University of Oxford; yasmine.elmasri@education.ox.ac.uk

Embretson, Susan; Georgia Institute of Technology; susan.embretson@psych.gatech.edu

Emery, Matthey; Roblox Corp; memery@roblox.com

Engelhard, George; UGA; gengelh@uga.edu

Englert, Kerry; Seneca Consulting, LLC; kenglert@senecaconsulting.org

Ercikan, Kadriye; Educational Testing Service; kercikan@ets.org

Erdemir, Aysenur; ; erdemiraysenur@gmail.com

Ersan, Ozge; University of Minnesota-Twin Cities; ersan001@umn.edu

Estabrooks, Carole A.; University of Alberta; cestabro@ualberta.ca

Evans, Carla M.; National Center for The Improvement of Educational Assessment; cevans@nciea.org

Evans, Josiah; Law School Admission Council; jevans@lsac.org

Everson, Howard; CUNY Graduate Center; howard.everson@gmail.com

Fager, Meghan; Hitachi Solutions America; megfager@gmail.com

Faiello, Matthew; University of Florida; matthew.faiello@ufl.edu

Falk, Carl; McGill University; carl.falk@mcgill.ca

Fan, Fen; NBME; FFan@nbme.org

Fan, Meng; ; mfan@humrro.org

Fan, Tingting; Nanjing University; tingtingfan121@hotmail.com

Fang, Yu; Law School Admission Council; yfang@lsac.org

Faulkner-Bond, Molly; WestEd; mfaulkn@wested.org

Fauss, Michael; ETS; mfauss@ets.org

Federiakin, Denis; Johannes Gutenberg University Mainz; denis.federiakin@gmail.com

Feiler, Jake; University of Alabama; jfeiler@crimson.ua.edu

Feinberg, Rich; National Board of Medical Examiners; rfeinberg@nbme.org

Feldberg, Zachary; University of Georgia; feldberg@uga.edu

Felline, Cosimo; PBS KIDS; cfelline@pbs.org

Feng, Tianying; UCLA; tfeng@cresst.org

Feng, Zechu; The University of Hong Kong; jason521@connect.hku.hk

Fernandez-Vina, Elizabeth A.; New Jersey Department of Education; lizfv817@gmail.com

Ferrara, Steve; HumRRO; sferrara1951@gmail.com

Feuerstahler, Leah; Fordham University; lfeuerstahler@fordham.edu

Fienberg, Michael; University of Southern California Rossier School of Education; fienberg@usc.edu

Participant Emails

(Last Name, First Name; Affiliation; Email)

Filonczuk, Audrey; University of Notre Dame; afiloncz@nd.edu
Fina, Anthony D.; University of Iowa; anthony-fina@uiowa.edu
Finch, Holmes; Ball State University; whfinch@bsu.edu
Fincher, Melissa; edCount, LLC; mfincher@edcount.com
Finn, Bridgid; Educational Testing Service; bfinn@ets.org
Finney, Sara; James Madison University; finneysj@jmu.edu
Firoozi, Tahereh; University of Alberta; tahereh@ualberta.ca
Fischbach, Antoine; University of Luxemburg; antoine.fischbach@uni.lu
Fishtein, Daniel; Educational Testing Service; DFISHTEIN@ets.org
Fitzpatrick, Steve; Pearson; s.fitzpatrick@pearson.com
Flor, Michael; Educational Testing Service; mflor@ets.org
Flowers, Tavia; ; t_flowers@hotmail.com
Flynn, Kylie; WestEd; kflynn2@wested.org
Fogle, Thomas; National Board of Medical Examiners; tfogle@nbme.org
Foley, Brett P.; Alpine Testing Solutions; brett.foley@alpinetesting.com
Forte, Ellen; edCount, LLC; eforte@edcount.com
Foster, David; Caveon Test Security; david.foster@caveon.com
Foster, Natalie; ; natalie.foster@oecd.org
Freeman, Jason; Georgia Institute of Technology; jason.freeman@gatech.edu
French, Brian; Washington State University; frenchb@wsu.edu
Frey, Bruce; University of Kansas; bfrey@ku.edu
Frey, Sharon; Riverside Insights; sharon.frey@riversideinsights.com
Fu, Jianbin; Educational Testing Service; jfu@ets.org
Fu, Yanyan; GMAC; frankyanyan@gmail.com
Fu, Yanyan; Graduate Management Admission Council; yfu@gmac.com
Fujimoto, Ken; Loyola University Chicago; kfujimoto@luc.edu
Furgol Castellano, Katherine; Educational Testing Service; KEcastellano@ets.org
Furter, Robert Thomas; Physician Assistant Education Association; rfurter@PAEAonline.org
Gagnon-Bartch, Johann; University of Michigan; johanngb@umich.edu
Galib, Linda; Loyola University Chicago; lgalib@luc.edu
Gamo, Sylvie; University of Luxemburg; sylvie.gamo@uni.lu
Gao, Furong; Human Resources Research Organization; furong_gao@yahoo.com
Gao, Rui; ETS; rgao@ets.org
Gao, Yizhu; ; yizhu@ualberta.ca
Garcia, Elda; National Association of Testing Professionals; elda.garcia@natponline.com
García-Minjares, Manuel; Universidad Nacional Autónoma de México (UNAM); manuel_garcia@cuaieed.unam.mx
Ge, Yuan; The College Board; yge@collegeboard.org
Geofroy, Stephen; The University of The West Indies; stephen.geofroy@sta.uwi.edu
Gerasimova, Daria; University of Kansas; gerasimova@ku.edu
Gershon, Richard; Northwestern University; gershon@northwestern.edu
Gianopulos, Garron; NWEA; garron.gianopulos@nwea.org
Gianopulos, Garron; NWEA; garron@gianopulos.com
Gierl, Mark; University of Alberta; mark.gierl@ualberta.ca
Gierl, Mark J; University of Alberta; mgierl@ualberta.ca
Gijbels, Liesbeth; University of Washington; lgijbels@uw.edu
Gilbar, Charlotte; Natrona County School District; charlotte_gilbar@natronaschools.org
Gocer Sahin, Sakine; New Meridian Corporation; sgocersahin@gmail.com
Gochyyev, Perman; University of California, Berk; perman@berkeley.edu
Godek, Ben; Cambium Assessment; benjamin.godek@cambiumassessment.com
Goh, Sugyung; Ewha Womans University; ksukyung15@ewhain.net
Goldhammer, Frank; DIPF | Leibniz Institute for Research and Information in Education, ZIB; goldhammer@dipf.de
Gong, Brian; Center for Assessment; bgong@nciea.org
González, Jorge; Pontificia Universidad Católica De Chile; jorge.gonzalez@mat.uc.cl
Gonzalez-Wegener, Xaviera; UCL Institute of Education; xgonzalezwe@gmail.com
Gooch, Reginald M; Educational Testing Services; rmgooch@ets.org
Goodman, Joshua; NCCPA; joshuag@nccpa.net
Goodrich, Shawna; Department of National Defence; shawnago@gmail.com
Goodwin, Amanda; Vanderbilt University; amanda.goodwin@vanderbilt.edu
Gorgun, Guher; University of Alberta; gorgun@ualberta.ca
Gorney, Kylie; University of Wisconsin-Madison; kyliengorney@gmail.com
Gotwals, Amelia; Michigan State University; gotwals@msu.edu
Grabovsky, Irina; National Board of Medical Examiners; igrabovsky@nbme.org
Graf, Edith Aurora; Educational Testing Service; agraf@ets.org
Green, Anson; Tyson Foods; Anson.Green@tyson.com
Greer, Eunice; Department of Education; eunice.greer@ed.gov
Greiff, Samuel; University of Luxembourg; samuel.greiff@uni.lu
Griger, Cassondra; University of Iowa; cgriger@uiowa.edu

Participant Emails

(Last Name, First Name; Affiliation; Email)

Griph, Gerald; Pearson; gerald.griph@pearson.com
Grochowalski, Joseph; College Board; joe.grochowalski@gmail.com
Gu, Dai; George Mason University; dgu4@gmu.edu
Gu, Lixiong; Educational Testing Service; lgu@ets.org
Gugiu, Mihaiela Ristei; National Registry of Emergency Medical Technicians; mgugiu@nremt.org
Guo, Hongwen; Educational Testing Service; hguo@ets.org
Guo, Wenjing; University of Alabama; wguo9@crimson.ua.edu
Guo, Yage; Center for Applied Linguistics; yguo@cal.org
Gwak, Yelin; ; youri7045@ewhain.net
Gyll, Sean; Western Governors University; sean.gyll@wgu.edu
Habermehl, Kyle; Pearson; kyle.habermehl@pearson.com
Habing, Brian; National Institute of Statistical Sciences; bhabing@niss.org
Hahnel, Carolin; DIPF | Leibniz Institute for Research and Information in Education, Centre for International Student; hahnel@dipf.de
Haines, R. Trent; Morgan State University; trent.haines@morgan.edu
Halpin, Peter; UNC-Chapel Hil; peter.halpin@unc.edu
Hamdani, Maria; Center for Measurement Justice; mhamdani@measurementjustice.org
Hamilton, Laura; American Institutes for Research; lhamilton@air.org
Han, Kahee; University of Kansas; kaheehan@ku.edu
Han, Kyung (Chris); Graduate Management Admission Council; khan@gmac.com
Han, Kyung (Chris) T.; Graduate Management Admission; truetheta@gmail.com
Han, Suhwa; University of Texas at Austin; suhwa@utexas.edu
Han, Youngjin; University of Maryland College Park; younghan@umd.edu
Han, Yuting; hanyuting716@gmail.com
Han, Yuting; Peking University; hanyuting@bjmu.edu.cn
Han, Zhuangzhuang; ETS; zhan@ets.org
Hansen, Mark; UCLA; markhansen@ucla.edu
Hao, Jiangang; Educational Testing Service; jhao@ets.org
Happel, Jay; College Board; jhappel@collegeboard.org
Harik, Polina; National Board of Medical Examiners; pharik@nbme.org
Haring, Samuel; ACT; samuel.haring@act.org
Harris, Deborah J; University of Iowa; debharris1027@gmail.com
Harry, Nisha; The University of The West Indies; nish2017h@gmail.com
Hathcoat, John D; James Madison University; hathcojd@jmu.edu
Hauger, Jeffrey; University of Massachusetts; jhauger@educ.umass.edu
Haviland, Sara; ETS; shaviland@ets.org
He, Qiwei; Educational Testing Service; qhe@ets.org
He, Surina; surina1@ualberta.ca
He, Yi; Edmentum; uiheyi@gmail.com
He, Yong; Measurement Incorporated; YHe@measinc.com
Heffernan, Neil T; Worcester Polytechnic Institute; nth@wpi.edu
Hellman, Scott; Pearson; scott.hellman@pearson.com
Hembry, Ian; MetaMetrics; ian.hembry@gmail.com
Hemphill, Cadelle; AIR; Chemphill@air.org
Hendrickson, Amy; College Board; ahendrickson@collegeboard.org
Henson, Robert; University of North Carolina at Greensboro; rahenson@uncg.edu
Heritage, Margaret; mheritag@ucla.edu
Hernandez, Diley; Georgia Tech; diley.hernandez@gatech.edu
Hernandez, Philip; philipah@stanford.edu
Hicks, Juanita; American Institutes for Research; juanita.hicks.11@gmail.com
Hille, Kathryn; ETS; khille@ets.org
Himelfarb, Igor; ihimelfarb@nbce.org
Hinds, Fiona; Independent Consultant; fiona@womeninmeasurement.org
Hirt, Ashley; ATLAS at the University of Kansas; a445h349@ku.edu
Ho, Andrew; Harvard Graduate School of Education; Andrew_Ho@gse.harvard.edu
Ho, Emily; emily-ho@northwestern.edu
Ho, Eric M; ericmho@g.ucla.edu
Ho, Jordan L; Altus Assessments; hojordan09@gmail.com
Ho, TsungHan; ETS; tho@ets.org
Hoben, Matthias; York University; mhoben@yorku.ca
Hodge, Kari; kari_hodge@alumni.baylor.edu
Hoffman, Alexander; AleDev Consulting; ahoffman@aledev.com
Höft, Lars; Leibniz Institute for Science and Mathematics Education; hoeft@leibniz-ipn.de
Hogan, Melissa; BrainPop; melissah@brainpop.com
Hojnoski, Robin; Lehigh University; roh206@lehigh.edu
Holtzman, Steven; Educational Testing Service; sholtzman@ets.org
Hong, Hyeri; California State University, Fresno; hyerihong@mail.fresnostate.edu
Hong, Minju; University of Arkansas; minjuh@uark.edu

Participant Emails

(Last Name, First Name; Affiliation; Email)

Hong, Yuan; ; yuan.hong@cambiumassessment.com
Hoover, Jeffrey; University of Kansas; jhoover4@ku.edu
Hopkins, David; Louisiana Department of Education; david.hopkins@la.gov
Hornung, Caroline; University of Luxemburg; caroline.hornung@ext.uni.lu
Howell, Heather; ETS; hhowell@ets.org
Howell, Jessica; The College Board; jhowell@collegeboard.org
Hu, Xiangen; University of Memphis; xiangenhu@gmail.com
Hua, Cheng; University of Alabama; chua@crimson.ua.edu
Huang, Chun-Wei; WestEd; chuang@wested.org
Huang, Mingya; mhuang233@wisc.edu
Huang, Qi; University of Wisconsin - Madison; qhuang85@wisc.edu
Huang, Sijia; Indiana University Bloomington; sijhuang@iu.edu
Hubert, Barbara; BrainPop; barbarah@brainpop.com
Hudson, Kim; Data Recognition Corporation; KHudson@datarecognitioncorp.com
Huff, Kristen; Curriculum Associates; khuff@cainc.com
Huff, Stacy R; ; srhuff@uncg.edu
Huggins-Manley, Anne Corinne; University of Florida; ahuggins@coe.ufl.edu
Huggins-Manley, Anne Corinne; University of Florida; amanley@coe.ufl.edu
Huh, NooRee; ACT, Inc; NooRee.Huh@act.org
Huo, Huade; AIR; hhuo@air.org
Hwu, Bo Sien;   BuBu130619@gmail.com
Ibanez Moreno, Beatriz; American Board of Surgery; bimoreno@absurgery.org
Ihlenfeldt, Samuel Dale; University of Minnesota; ihlen010@umn.edu
Ing, Marsha; University of California, Rive; marsha.ing@ucr.edu
Ingrisone, James; Pearson VUE; ingrisone@gmail.com
Ingrisone, Soo; Pearson; singrisone@gmail.com
Inostroza Fernández, Pamela Isabel; University of Luxemburg; pamela.inostroza@uni.lu
Ivanova, Militsa; University of Cyprus; militsagi@yahoo.com
Jackson, Janine; Morgan State University; Jajac64@morgan.edu
Jackson, Kara; University of Washington; karajack@uw.edu
Jackson, Yvette Yvette; ; yvlee1@morgan.edu
Jafari, Amir; Cambium Assessment; amir.jafari@cambiumassessment.com
Jansen, Thorben; Leibniz Institute for Science and Mathematics Education; tjansen@leibniz-ipn.de
Jeffries, Jay; University of Nebraska-Lincoln; jayjeffries13@huskers.unl.edu
Jeon, Eunjeong; Ewha Womans University; euneun1202@ewhain.net
Jeon, Minjeong; UCLA; mjjeon@ucla.edu
Jeong, Hyo; Sogang University; hshinedu@sogang.ac.kr
Jeppson, Haley; National Institute of Statistical Sciences; hjeppson@niss.org
Jewsbury, Paul Adrian; Educational Testing Service; pjewsbury@ets.org
Ji, Feng; ; fengji@berkeley.edu
Ji, Xuejun Ryan; The University of British Columbia; x.ryan.ji@gmail.com
Jia, Hao; National Council of State Boards of Nursing (NCSBN); hjia@ncsbn.org
Jia, Yue; Educational Testing Service; yjia@ets.org
Jiang, Chunlian; University of Macau; cljiang@um.edu.mo
Jiang, Ning; ; jiangning0302@gmail.com
Jiang, Tao; Cambium Assessment, Inc; tao.jiang@cambiumassessment.com
Jiang, Yang; ETS; yjiang002@ets.org
Jiang, Zhehan; ; jiangzhehan@gmail.com
Jiang, Zhehan; ; zjiang4@outlook.com
Jiao, Hong; University of Maryland; hjiao@umd.edu
Jin, Kuan Yu; ; kyjin@hkeaa.edu.hk
Jin, Yi; ; kimmm@connect.hku.hk
Jin, Ying; Association of American Medical Colleges; yjin@aamc.org
Johnson, Hayden; MN; joh18320@umn.edu
Johnson, Janice Lee; NWEA; janice.johnson@nwea.org
Johnson, Mark; Cognia, Inc.; mark.johnson@cognia.org
Johnson, Matthew; ETS; msjohnson@ets.org
Jones, Andrew; American Board of Surgery; ajones@absurgery.org
Jones, Paul Edward; Pearson VUE; paul.jones@pearson.com
Jonson, Jessica L.; Buros Center for Testing-UNL; jjonson2@unl.edu
Joo, Sean; University of Kansas; sjoo@ku.edu
Jorgensen, Terrence D; University of Amsterdam; T.D.Jorgensen@uva.nl
Jozkowski, Kristen; Indiana University; knjozkow@iu.edu
Julian, Marc W; DRC; mjulian@datarecognitioncorp.com
Jung, Hyun Joo; University of Massachusetts Amherst; hyunjoo.jung2@gmail.com
Jung, Ji Yoon; Boston College; jiyoon.jung@bc.edu
Jung, Juyoung; The University of Iowa; juyoung-jung@uiowa.edu

Participant Emails

(Last Name, First Name; Affiliation; Email)

Jurich, Daniel; National Board of Medical Examiners; DJurich@nbme.org
Kaliski, Pamela; ABIM; PKaliski@ABIM.ORG
Kamata, Akhito; Southern Methodist University; akamata@smu.edu
Kamata, Akihito; Southern Methodist University; akamata@gmail.com
Kane, Michael; Educational Testing Service; mkane@ets.org
Kang, Hyeon-Ah; University of Texas at Austin; hkang@austin.utexas.edu
Kannan, Priya; WestEd; pkannan@wested.org
Kanopka, Klint; ; kkanopka@stanford.edu
Kao, Shu-chuan; NCSBN; skao@ncsbn.org
Kaplan, David; University of Wisconsin - Madison; david.kaplan@wisc.edu
Kapoor, Radhika; ; rkap786@stanford.edu
Kapoor, Shalini; ACT, Inc.; shalinikapoor.ia@gmail.com
Kara, Yusuf; Southern Methodist University; ykara@smu.edu
Karamese, Hacer; WIDA at University of Wisconsin; hacer.karamese@wisc.edu
Karamese, Hacer; WIDA at University of Wisconsin-Madison; karamese@wisc.edu
Kares, Faith R; Beloved Community; faith@wearebeloved.org
Kartal, Gamze; University of Illinois at Urbana-Champaign; gkartal2@illinois.edu
Katz, Daniel; NWEA; daniel.katz@nwea.org
Kayton, Heather Leigh; University of Oxford; heather.kayton@education.ox.ac.uk
Keefe, Janice; Mount Saint Vincent University; Janice.Keefe@msvu.ca
Kehinde, Olasunkanmi; Washington State University; kehinde.james@wsu.edu
Kelberlau, Darin; Millard Public Schools; dckelberlau@mpsomaha.org
Kell, Harrison; ETS; hkell@ets.org
Keller, Lisa; University of Massachusetts; lkeller@educ.umass.edu
Keller-Margulis, Milena A.; University of Houston; mkmargulis@Central.UH.EDU
Kendall Brooks, Lauren; Advanced Education Research and Development Fund; lkendallbrooks@aerdf.org
Keng, Leslie; Center for Assessment; lkeng@nciea.org
Kennedy, Patrick; University of Oregon; ppaine@uoregon.edu
Kern, Justin L.; University of Illinois at Urbana-Champaign; kern4@illinois.edu
Kerzabi, Emily; Educational Testing Service; ekerzabi@ets.org
Ketterlin Geller, Leanne; Southern Methodist University; lkgeller@mail.smu.edu
Ketterlin Geller, Leanne; Southern Methodist University; lkgeller@smu.edu
Khorramdel, Lale; Boston College; lale.khorramdel@bc.edu
Kilmen, Sevilay ; kilmen@ualberta.ca
Kilmen, Sevilay; Bolu Abant Izzet Baysal University; sevilaykilmen@gmail.com
Kim, Ahyoung; WIDA, University of Wisconsin; alicia.a.kim@gmail.com
Kim, Brian H; Commonapp Inc.; bkim@commonapp.org
Kim, Dong-In; Data Recognition Corporation; DKim@DataRecognitionCorp.com
Kim, Hyung Jin; University of Iowa; hyungjin-kim@uiowa.edu
Kim, Ji-Hye; Korean Educational Development Institute; jihyekim@kedi.re.kr
Kim, JongPil; Riverside Insights; jp.kim@riversideinsights.com
Kim, Kyung Yong; University of North Carolina Greenboro; k_kim9@uncg.edu
Kim, Minsung; ACT, Inc.; doug.kim@act.org
Kim, Seongeun; University of North Carolina at Greensboro; s_kim45@uncg.edu
Kim, Stella; University of North Carolina at Charlotte; stella-kim@uncc.edu
Kim, Sujin; George Mason University; skim222@gmu.edu
Kim, Sungyeun; Incheon National University; syk@inu.ac.kr
Kim, Sunhee; ; sunnyk0206@yahoo.com
Kim, Sunhee; College Board; sunkim@collegeboard.org
Kim, Woomee; George Mason University; wkim18@gmu.edu
Kim, Young Yee; American Institutes for Research; ykim@air.org
Kim, YoungKoung; College Board; ykim@collegeboard.org
Kim, Youngwon; University of Washington; kimyw@uw.edu
Kim, Yun-Kyung; UCLA; yunkim2729@ucla.edu
King, Sarah; University of Texas at Austin; sarah.gorsky@utexas.edu
Kingsbury, G. Gage; ; gagekingsbury@comcast.net
Kinsey, Devon; Educational Testing Service; DKINSEY001@ets.org
Kintz, Tara; Michigan Assessment Consortium; kintztar@msu.ed
Klieger, David; ETS; dklieger@ets.org
Knight, Melondy; Curriculum Associates; mknight@cainc.com
Ko, Amy; University of Washington; ajko@uw.edu
Kobrin, Jennifer; ATLAS: University of Kansas; jennifer.kobrin@ku.edu
Koehn, Hans Friedrich; ; hkoehn@illinois.edu
Kolen, Michael; The University of Iowa; kolenmichael@gmail.com
Kosh, Audra; NWEA; audrakosh@gmail.com
Koval, Jayma; Georgia Institute of Technology; jayma.koval@ceismc.gatech.edu
Kozikowski, Andrzej; NCCPA; andrzejk@nccpa.net

Participant Emails

(Last Name, First Name; Affiliation; Email)

Kroehne, Ulf; ; kroehne@dipf.de
Kronberg, Carla; UWI St Augustine; carla.kronberg@sta.edu.uwi
Kroopnick, Marc; Association of American Medical Colleges; mkroopnick@aamc.org
Krost, Kevin; Virginia Polytechnic Institute and State University; kevinkrost@vt.edu
Kuhfeld, Megan; NWEA; megan.kuhfeld@nwea.org
Kukea Shultz, Pohai; University of Hawaii at Manoa; pohai@hawaii.edu
Kuklick, Livia; IPN Kiel, Germany; kuklick@leibniz-ipn.de
Kumar, Lavanya Shravan; University of South Florida; lkumar1@usf.edu
Kunina-Habenicht, Olga; ; olga.kunina-habenicht@tu-dortmund.de
Kuo, Tzu-Chun; Kaplan North America; tzu710088@gmail.com
Kuzinets, Vladimir; NBOME; VKuzinets@nbome.org
Kwako, Alexander; ; akwako@ucla.edu
Kwon, Sunbeom; ; sunbeom2@illinois.edu
Kyllonen, Patrick Charles; ETS; pkyllonen@ets.org
La Torre, Deborah; UCLA; latorre@cresst.org
LaFlair, Geoff; Duolingo; geoff@duolingo.com
Lai, Hollis; ; hollis1@ualberta.ca
Lai, Ka Wing; ; Kawing2@ualberta.ca
Lai, Viet; University of Oregon; vietl@uoregon.edu
Laitusis, Cara Cahalan; ETS; claitusis@ets.org
Lakin, Joni; University of Alabama; Jlakin@ua.edu
Lam, Jenny; University of Alberta; jlam3@ualberta.ca
Lan, Andrew; University of Massachusetts at Amherst; andrewlan@cs.umass.edu
Landl, Erika; Center for Assessment; ehall@nciea.org
Lane, John; ; lanejoh3@msu.edu
Lane, Suzanne; University of Pittsburgh; sl@pitt.edu
lang, david; Stanford University; dlang@stanford.edu
Langi, Meredith; NWEA; meredith.langi@nwea.org
Larson, Eric C.; Southern Methodist University; eclarson@smu.edu
Laster, Christina; National Parents Union; christina@npunion.org
Laukaityte, Inga; Umeå University; inga.laukaityte@umu.se
Law, Evelyn; National University of Singapore; paelecn@nus.edu.sg
Law, Nancy; University of Hong Kong; nlaw@hku.hk
Lawless, Rene; ETS; rlawless@ets.org
Lazarus, Sheryl; National Center On Educational; laza0019@umn.edu
LeBeau, Brandon; University of Iowa; brandon-lebeau@uiowa.edu
Lee Brown, Taneisha; The Findings Group; taneisha@thefindingsgroup.org
Lee, Boram; Ewha Womans University; boram.lee@ewha.ac.kr
Lee, Chansoon; American Board of Internal Medicine; dlee@abim.org
Lee, Daniel Yangsup; College Board; yangsupl2@gmail.com
Lee, Dayeon; ; from.dayeon@gmail.com
Lee, Dukjae; University of Massachusetts Am; dlee@umass.edu
Lee, Eunji; George Washington University; ejlee7787@gwu.edu
Lee, Haeju; University of North Carolina Greenboro; HLEE@uncg.edu
Lee, Hollylynne; North Carolina State University; hstohl@ncsu.edu
Lee, Hyeryung; ; hyerlee@uiowa.edu
Lee, Hyunjung; ; hlee201@fordham.edu
Lee, Hyunsook; ; yhsishappy0818@gmail.com
Lee, Jade Caines; University of Kansas; jade.caines.lee@gmail.com
Lee, Minho; University of California Los Angeles; leemino72@ucla.edu
Lee, Sehee; ; 2016110555@swu.ac.kr
Lee, Seo Young; Prometric LLC; seoyoung.lee@prometric.com
Lee, Sooyong; University of Texas; sooyongl09@utexas.edu
Lee, Sunhyoung; University of Nebraska-Lincoln; slee82@huskers.unl.edu
Lee, Won-Chan; University of Iowa; won-chan-lee@uiowa.edu
Lee, Yi-Hsuan; Educational Testing Service; ylee@ets.org
Lee, Yongseok; University of Florida; yseok.lee@ufl.edu
Lee, Yoonsun; Seoul Women's University; ylee@swu.ac.kr
Leeming, Meghan; University of North Carolina at Greensboro; mkleemin@uncg.edu
Lehrfeld, Jon; Educational Testing Service; jlehrfeld@ets.org
Leino, Rosa; Standards & Testing Agency; Rosa.leino@education.gov.uk
Lembke, Erica; University of Missouri; lembkee@missouri.edu
Leng, Dihao; Boston College; dihao.leng@bc.edu
Lentini, Jennifer; Educational Testing Service (ETS); jlentini@ets.org
Leonard, Peter; Chicago Public Schools; pjleonard1@cps.edu
Leung-Gagné, Josh; Stanford University; jgagne@stanford.edu
Leventhal, Brian C; James Madison University; leventbc@jmu.edu

Participant Emails

(Last Name, First Name; Affiliation; Email)

Levine Brown, Elizabeth; George Mason University; ebrown11@gmu.edu
Levine, Felice; American Educational Research Association; flevine@aera.net
Levy, Roy; Arizona State University; roy.levy@asu.edu
Lewis, Daniel; Creative Measurement Solutions LLC; dan.lewis@creativemeasurement.com
Lewis, Jennifer L.; University of Massachusetts Am; jlewi0@umass.edu
Li, Chen; ETS; cli@ets.org
Li, Guiyu; East China Normal University; guiyuli@outlook.com
Li, Isaac; ; liy1@mail.usf.edu
Li, Jie; Ascend Learning; lijdbc@gmail.com
Li, Jie; NCS - Pearson; jie.li@pearson.com
Li, Jiehan; ; jiehanli@ufl.edu
Li, Lanrong; Amplify Education; lli@amplify.com
Li, Lanrong; Amplify Education, Inc.; jessicalilr2011@gmail.com
Li, Linlin; WestEd; lli@wested.org
Li, Min; University of Washington; minli@u.washington.edu
Li, Shuhong; ETS; sli@ets.org
Li, Tongyun; Educational Testing Service; tli002@ets.org
Li, Wen-Ching; Data Recognition Corporation; WLi@datarecognitioncorp.com
Li, Xin; ACT, Inc.; xin.li@act.org
Li, Xin (Grace); University of Wisconsin-Madison; gracexinlee@gmail.com
Li, Xueming; NWEA; sylvia.li@nwea.org
Li, Yalin; Beijing Insight Online Management Consulting Co.,Ltd; lyl199681@aliyun.com
Li, Zhen; eMetric LLC; zli@emetric.net
Li, Zhushan Mandy; Boston College; zhushan.li@bc.edu
Li, Ziying; University of Florida; ziying.li@ufl.edu
Li, Zonggui; Boston College; il@bc.edu
Lian, Xu; Beijing Insight Online Management Consulting Co.,Ltd; lianx@zhiding.com.cn
Liang, Min; ; min-liang-1@uiowa.edu
Liang, Qianru; The University of Hong Kong; liangqr@hku.hk
Liang, Xinya; Department of Rehabilitation, Human Resources and Communication Disorders, University of Arkansas; xl014@uark.edu
Liao, Dandan; Mckinsey & Company; dandanliao0518@gmail.com
Liao, Manqian; Duolingo; mancy@duolingo.com
Liao, Xiangyi; ; xliao36@wisc.edu
Liaw, Yuan-Ling; ; yuan-ling.liaw@iea-hamburg.de
Lim, Hwanggyu; Graduate Management Admission Council; hglim83@gmail.com
Lim, Jared; University of Washington; jorlim7@uw.edu
Lim, Sangdon; University of Texas at Austin; sangdonlim@utexas.edu
Lim, Youn Seon; Univeristy of Cincinnati; limyo@ucmail.uc.edu
Lin, Haiyan; Pearson; Haiyan.lin@pearson.com
Lin, Jia; Howard University; jialin1984@gmail.com
Lin, Ye; Ascend Learning; yelin.nora@gmail.com
Lin, Zhongtian; Cambium Assessment, Inc; zhongtian.lin@cambiumassessment.com
Lindner, Marlit; IPN Kiel, Germany; mlindner@leibniz-ipn.de
Lindner, Marlit Annalena; IPN Kiel; mlindner@ipn.uni-kiel.de
Lindsay, Constance; University of North Carolina at Chapel Hill; clindsay@unc.edu
Ling, Guangming; Educational Testing Service; gling@ets.org
Lissitz, Robert; University of Maryland; rlissitz@umd.edu
Liu, Allison; Worcester Polytechnic Institute; aliu2@wpi.edu
Liu, Chunyan; National Board of Medical Examiners; cliu@nbme.org
Liu, Guangyun; ; lguangyun@uiowa.edu
Liu, Huan; The University of Iowa; huan-liu-1@uiowa.edu
Liu, Jingchen; Columbia University; jcliu@stat.columbia.edu
Liu, Jinghua; Pearson; jinghua.liu@pearson.com
Liu, Jinghua; Pearson; jinghuawalkerliu@gmail.com
Liu, Juan; Beijing Insight Online Management Consulting Co.,Ltd; liuj_psy@outlook.com
Liu, Lei; Educational Testing Service; lliu001@ets.org
Liu, Liu; University of Washington; liuuil@uw.edu
Liu, Lucia; Ascend Learning; Lucy.Xin.Liu@gmail.com
Liu, Lydia T.; Cornell University; lydiatliu@berkeley.edu
Liu, Ou Lydia; ETS; lliu@ets.org
Liu, Ren; University of California, Merc; rliu45@ucmerced.edu
Liu, Xiang; Educational Testing Service; xliu003@ets.org
Liu, Xiangdong; University of Iowa; xiangdong-liu@uiowa.edu
Liu, Xin Lucy; Ascend Learning; Xin.Liu@ascendlearning.com
Liu, Yang; University of Maryland, College Park; yliu87@umd.edu
Liu, Ying; University of Illinois at Urbana-Champaign; yingl7@illinois.edu

Participant Emails

(Last Name, First Name; Affiliation; Email)

Livne, Oren; Educational Testing Service; olivne@ets.org
Lo, Wen-Juo; Unversity of Arkansas; wlo@uark.edu
Locke, Victoria; Istation; vlocke@istation.com
Lockwood, J.R.; Duolingo; jr@duolingo.com
Loken, Eric Eric; University of Connecticut; eric.loken@uconn.edu
Longe, Brendan; University of Massachusetts Amherst; blonge@umass.edu
Lopera-Oquendo, Carolina; CUNY Graduate Center; cloperaoquendo@gradcenter.cuny.edu.co
Lorie, William A.; Center for Assessment; william.lorie@gmail.com
Lottridge, Susan; Cambium Assessment; susan.lottridge@cambiumasssessment.com
Lottridge, Susan; Cambium Assessment, Inc; susan.lottridge@cambiumassessment.com
Love, Quentin Ulysses Adrian; WestEd; quintinulove@gmail.com
Love, Quintin; New Meridian Corporation; qlove@newmeridiancorp.org
Lovelace, Temple S; Advanced Education Research and Development Fund; templelovelace@gmail.com
Lu, Ru; Educational Testing Service; rlu@ets.org
Lu, Yang; Pearson; Yang.Lu@pearson.com
Lu, Yikai; University of Notre Dame; ylu22@nd.edu
Lu, Ying; College Board; ylu@collegeboard.org
Lucas, Tracey Michelle; University of The West Indies; tracey.m.lucas@gmail.com
Luecht, Richard Melvin; University of North Carolina at Greensboro; rmluecht@uncg.edu
Lugu, Benjamin Kweku; ; benjamin.lugu@aims.ac.rw
Luo, Jinwen; UCLA; jevan.luo@gmail.com
Luo, Yachen; ; yl15j@my.fsu.edu
Luo, Yong; NWEA; jackyluoyong@gmail.com
Lyons, Susan; Lyons Assessment Consulting; susan@lyonsassessment.com
Lyu, Weicong; University of Wisconsin - Madison; wlyu4@wisc.edu
Ma, Chenchen; University of Michigan; chenchma@umich.edu
Ma, Cheryl; Amazon Web Services (AMS); cherylyema@gmail.com
Ma, Jing; The University of Iowa; majing@uiowa.edu
Ma, Wanjing Anya; Stanford University; wanjingm@stanford.edu
Ma, Wenchao; University of Alabama; wenchao.ma@ua.edu
Ma, Ye; AWS; ymcheryl@amazon.com
MacGregor, David; Center for Applied Linguistics; dmacgregor@cal.org
MacIntosh, Alexander; Altus Assessments; amacintosh@altusassessments.com
Madison, Matthew James; University of Georgia; mjmadison@uga.edu
Maeda, Hotaka; Smarter Balanced; hotaka.maeda@gmail.com
Maeda, Yukiko; Purdue University; ymaeda@purdue.edu
Maibach, Jacob; University of Arizona; jmaibach@arizona.edu
Malatesta, Jaime; Graduate Management Admission Council; jmalatesta@gmac.com
Man, Kaiwen; University of Alabama; mankaiwen@hotmail.com
Mangino, Tony A; University of Kentucky; Anthony.Mangino@uky.edu
Manna, Venessa; Educational Testing Service; vmanna@ets.org
Mao, Xia; NBOME; xmao@nbome.org
Mardones, Constanza; University of Georgia; cam04214@uga.edu
Marion, Scott; National Center for The Improvement of Educational Assessment; smarion@nciea.org
Martin de los Santos Kleinz, Lisa; Johannes Gutenberg University of Mainz; limartin@uni-mainz.de
Martineau, Joseph A.; Educational Testing Service; jmartineau@ets.org
Martinez, Alfonso; University of Iowa; alfonso-martinez@uiowa.edu
Martinez, Jose Felipe; UCLA - School of Education and Information Studies; jfmtz@ucla.edu
Martínez-González, Adrián; Universidad Nacional Autónoma de México (UNAM); adrian_martinez@cuaieed.unam.mx
Martin-Raugh, Michelle; University of Texas Arlington; michelle.martinraugh@uta.edu
Matta, Michael; University of Houston; mmatta@uh.edu
Matta, Tyler; NWEA; tyler.matta@nwea.org
Mattison, Kate; IXL Learning; kmattison@ixl.com
Maur, Andreas; Johannes Gutenberg University of Mainz; anmaur@uni-mainz.de
Mawhinney, Lynnette; Rutgers-Newark; Lynnette.mawhinney@rutgers.edu
Maxwell, Liam James; Standards & Testing Agency; liam.maxwell@education.gov.uk
Mbelede, Njideka Gertrude; Nnamdi Azikiwe University Awka; ng.mbelede@unizik.edu.ng
McCaffrey, Daniel; Educational Testing Service; dmccaffrey@ets.org
McCall, Martha; Mckinsey & Company; mccall.marty@gmail.com
Mcclure, Kyla; University of Colorado Boulder; Kyla.Mcclure@colorado.edu
McCormick, Carina M.; University of Nebraska-Lincoln; carinamc@gmail.com
McCullough, Janeen; Law School Admission Council; mccullough.janeen@gmail.com
McFadden, Mara; James Madison University; mcfad2me@jmu.edu
McGrane, Joshua; University of Oxford; joshua.mcgrane@education.ox.ac.uk
McHugh, Bridget; Center On Education and Training for Employment; mchugh.159@osu.edu
McKlin, Tom; The Findings Group; tom@thefindingsgroup.org
McMahon, Margaret; Cambium Assessment; margaret.mcmahon@cambiumassessment.com
McMurtry, Teaira; University of Alabama, Birmingham; mcmurtry@uab.edu

Participant Emails

(Last Name, First Name; Affiliation; Email)

McNamara, Danielle; Arizona State University; dsmcnamara@gmail.com
McNeish, Daniel; Arizona State University; dmcneish@asu.edu
McVay, Aaron; Law School Admission Council; amcvay@lsac.org
Medina Morales, Norma; Amplify Education; nmedinamorales@amplify.com
Mee, Janet; NBME; jmee@nbme.org
Melaco, Carla; Beloved Community; carla@wearebeloved.org
Mendiola, Melchor Sanchez; Universidad Nacional Autonoma de Mexico (UNAM); melchorsm@unam.mx
Meng, Huijuan; Amazon Web Services (AWS); huijuam@amazon.com
Mercer, Sterett H.; The University of British Columbia; sterett.mercer@ubc.ca
Merkle, Edgar; ; merklee@missouri.edu
Merriman, Jennifer; International Baccalaureate; jen.merriman@ibo.org
Mertens, Ute; Leibniz Institute for Science and Mathematics Education; mertens@leibniz-ipn.de
Meyer, Jennifer; Leibniz Institute for Science and Mathematics Education; jmeyer@leibniz-ipn.de
Miao, Jing; Educational Testing Service; jmiao@ets.org
Michaelides, Michalis; ; michaelides.michalis@ucy.ac.cy
Michel, Rochelle; Smarter Balanced; rochelle.michel@gmail.com
Michels, Michael Andreas; ; michael.michels@uni.lu
Middlestead, Andrew J.; Michigan Department of Education; middlesteada@michigan.gov
Middleton, Kyndra; Howard University; kvmiddleton@gmail.com
Migunov, Igor; Fordham University; imigunov@fordham.edu
Mikeska, Jamie; ETS; Jmikeska@ets.org
Miller, Angela; George Mason University; amille35@gmu.edu
Miller, M. David; University of Florida; dmiller@coe.ufl.edu
Miller, Sherral; College Board; shmiller@collegeboard.org
Milligan, Sandra; The University of Melbourne; s.milligan@unimelb.edu.au
Mills, Christine; Ascend Learning;
Miranda, Alejandra; ; miran143@umn.edu
Mislevy, Robert J.; Retired; rmislevy@umd.edu
Missall, Kristen; University of Washington; kmissall@uw.edu
Mitchell, Jamie; Stanford University; jamiel12@stanford.edu
Mo, Ya; Boise State; yamo@boisestate.edu
Mohammadi, Hamid; Department of Computer Engineering, University of Amirkabir; hamid.mohammadi@aut.ac.ir
Mojoyinola, Mubarak Olumide; The University of Iowa; mubarak-mojoyinola@uiowa.edu
Moncaleano, Sebastian; Curriculum Associates, LLC; seb.moncaleano@gmail.com
Monroe, Scott; UMass Amherst; smonroe@educ.umass.edu
Montgomery, Melinda; Pearson; melindasmontgomery@gmail.com
Moore, Allen Christopher; University of Georgia; acm01415@uga.edu
Moreno Luna, Arlyn Y; University of California, Berkeley; arlyn_morenoluna@berkeley.edu
Morrell, Monica; University of Maryland; m.morell003@gmail.com
Morrison, Kristin M.; Curriculum Associates; KMorrison@cainc.com
Moses, Tim; College Board; tmoses@collegeboard.org
Moskowitz, Joshua; Altus Assessments; jmoskowitz@altusassessments.com
Moustaki, Irini; London School of Economics and Political Science; i.moustaki@lse.ac.uk
Muckle, Timothy; NBCRNA; tmuckle@nbcrna.com
Mueller, Lorin; Federation of State Boards of Physical Therapy; lmueller@fsbpt.org
Muenzen, Patricia; ACT; patricia.muenzen@act.org
Mugan, Leslie; NWEA; leslie.mugan@nwea.org
Mulvihill, Megan M; University of Kansas; megan.mulvihill@ku.edu
Muniz, Jose; University of Nebrija; jmuniz@nebrija.es
Muniz, Jose; University of Oviedo; jmuniz@uniovi.es
Munson, Liberty; Microsoft; Liberty.Munson@microsoft.com
Muntean, William J; National Council of State Boards of Nursing; williamjmuntean@gmail.com
Murphy, Daniel; WestEd; dmurphy@wested.org
Murphy, Victoria A; University of Oxford; victoria.murphy@education.ox.ac.uk
Myers, Aaron; American Board of Internal Medicine; amyers@abim.org
Myers, Matthew; ; mcmyers@udel.edu
Nabors Olah, Leslie; Educational Testing Service; lnaborsolah@ets.org
Nagel, Marie-Theres; Johannes Gutenberg University of Mainz; marie.nagel@uni-mainz.de
Nájera, Pablo; Autonomous University of Madrid; pablo.najera@uam.es
Namsone, Dace; University of Latvia; dace.namsone@lu.lv
Nascimento, Wallace; University of Florida; wallacenpj@gmail.com
Naveiras, Matthew David; Peabody College of Vanderbilt; matthew.d.naveiras@vanderbilt.edu
Nemeth, Yvette M; HumRRO; ynemeth@humrro.org
Nesbitt, Jaylin N; James Madison University; nesbitjn@dukes.jmu.edu
Nese, Joseph F. T.; University of Oregon; jnese@uoregon.edu
Nguyen, Thien; University of Oregon; thien@cs.uoregon.edu
Niazi, Farhan; University of Iowa; farhan-niazi@uiowa.edu

Participant Emails
(Last Name, First Name; Affiliation; Email)

Nickodem, Kyle; University of North Carolina - Chapel Hill; knickodem@unc.edu
Nie, Chang; Beijing Normal University; niechang@mail.bnu.edu.cn
Nolan, Katherine; Curriculum Associates; NolanKatherineA@gmail.com
Nydick, Steven; Duolingo; steven@duolingo.com
Nye, Christopher; Michigan State University; nyechris@msu.edu
Ober, Teresa; Educational Testing Service (ETS); teresaober@gmail.com
O'Donnell, Francis; National Board of Medical Examiners; fodonnell@nbme.org
O'Dwyer, Eowyn; Educational Testing Service; eowyer@ets.org
Ogut, Burhan; American Institutes for Research; bogut@air.org
Oh, Hyeon-Joo; Riverside Insights; joannehj@gmail.com
Oh, Kyuseol; Geunhwa Girls High School; kyuseol@gmail.com
Olgar, Suleyman; Florida Department of Education; suleyman.olgar@fldoe.org
Olivera Aguilar, Margarita; Educational Testing Service; molivera-aguilar@ets.org
Oliveri, Maria Elena; Buros Center for Testing-UNL; oliveri.m@live.com
Oliveri, Maria Elena; University of Nebraska Lincoln; moliveri@buros.org
Oliveri, Maria-Elena; University of Nebraska; moliveri2@unl.edu
Oluwalana, Olasumbo; Educational Testing Service; ooluwalana@ets.org
O'Neil, Harold; ; honeil@usc.edu
O'Neill, Richard; Peoria Unified School District; roneill@pusd11.net
O'Neill, Thomas; ABFM; toneill@theabfm.org
Ong, Thai Quang; National Board of Medical Examiners; tong@nbme.org
Onna, Marieke van; Cito; marieke.vanonna@cito.nl
O'Reilly, Tenaha; Educational Testing Service; toreilly@ets.org
Ormerod, Christopher; Cambium Assessment; christopher.ormerod@cambiumassessment.com
O'Rourke, Hannah; University of Alberta; hannah.orourke@ualberta.ca
O'Rourke, Kevin P.; ; kporourke@umass.edu
Ottmar, Erin; Worcester Polytechnic Institute; erottmar@wpi.edu
Ouyang, Jing; University of Michigan; jingoy@umich.edu
Ouyang, Jinying; ; ouyangjinying@m.scnu.edu.cn
Ouyang, Wenli; National Board of Medical Examiners; wouyang@nbme.org
Ozkeskin, Emrah Emre; Heilbronn University, Germany; emre.ozkeskin@gmail.com
Paccagnella, Marco; OECD; Marco.PACCAGNELLA@oecd.org
Padgett, R Noah; Baylor University; Noah_Padgett1@baylor.edu
Padilla, Geraldo Bladimir; ; gbpadilla@uiowa.edu
Padro Collazo, Pascua; University of Puerto Rico; pascua.padro@upr.edu
Palermo, Corey; Measurement Incorporated; cpalermo@measinc.com
Palma, Jose R.; The University of Texas at Austin; jose.palma@austin.utexas.edu
Pan, Qianqian; National Institute of Education, Nanyang Technological University; panqianqian2013@gmail.com
Pan, Tianshu; Pearson; tianshu.pan@pearson.com
Pan, Yiqin; University of Florida; ypan@coe.ufl.edu
Papanastasiou, Elena; University of Nicosia; papanastasiou.e@unic.ac.cy
Pappas, Sandra; Amplify Education; spappas@amplify.com
Paris, Joseph; West Chester University; jparis@temple.edu
Park, Elizabeth; Educational Testing Service; expark@ets.org
Park, Jung Yeon; George Mason University; jpark233@gmu.edu
Park, Seohee; American Board of Internal Medicine; spark@abim.org
Parks, Charles; CRESST/UCLA; cparks@cresst.org
Parra-Martinez, Fabio Andres; University of Arkansas; ap448@uark.edu
Patarapichayatham, Chalie; NWEA; chalie.patara@nwea.org
Patelis, Thanos; Johns Hopkins U & University of Kansas; tpatelis@yahoo.com
Patterson, Chris; James Madison University; patte3cr@dukes.jmu.edu
Patterson, Jim; College Board; jpatterson@collegeboard.org
Patterson, Luke; AIR; lpatterson@air.org
Patz, Richard; UC Berkeley; rpatz@berkeley.edu
Peabody, Michael R; National Association of Boards of Pharmacy; michael.peabody77@gmail.com
Pei, Bo; University of Notre Dame; bpei@nd.edu
Pellegrino, James; University of Illinois at Chicago; pellegjw@uic.edu
Perez, Alexandra Lane; University of Connecticut; alexandra.stone@uconn.edu
Perez, Joselyn; ; joselyn.perez@uconn.edu
Perie, Marianne; WestEd; mp@measurementinpractice.com
Pham, Duy N.; Educational Testing Service; dnpham@ets.org
Pham, Paul; University of Washington; pkdpham@uw.edu
Phenow, Aurore; Data Recognition Corporation; APhenow@datarecognitioncorp.com
Phenow, Aurore Yang; Data Recognition Corporation; aurore.phenow@gmail.com
Phillip, kristy; The University of The West Indies; kristy.phillip@my.uwi.edu
Phillip, Sharon; The University of The West Indies; sharon.phillip@sta.uwi.edu
Phillippo, Kate; Loyola University Chicago; kphillippo@luc.edu

Participant Emails

(Last Name, First Name; Affiliation; Email)

Piacentini, Mario; OECD; mario.piacentini@oecd.org
Pitoniak, Mary; ETS; mpitoniak@ets.org
Plackner, Christie; Data Recognition Corporation; cplackner@datarecognitioncorp.com
Pohl, Steffi; Freie Universitat Berlin; steffi.pohl@fu-berlin.de
Por, Han-Hui; ETS; hpor@ets.org
Potgieter, Cornelis; Texas Christian University; c.potgieter@tcu.edu
Potter, Andrew; University of Delaware; ahpotter@udel.edu
Powers, Sonya; Edmentum, Inc.; sonya.powers@edmentum.com
Powers, Sonya; Edmentum, Inc.; sopowers@gmail.com
Prendez, Jordan Yee; American Board of Internal Medicine; jordanyeeprendez@protonmail.com
Price, Argenta; Stanford University; argenta@stanford.edu
Prihar, Ethan B; Worcester Polytechnic Institute; ebprihar@wpi.edu
Proctor, Thomas; College Board; tproctor@collegeboard.org
Puhan, Gautam; ETS; gpuhan@ets.org
Purpura, David; Purdue University; purpura@purdue.edu
Qazi, Nabeel; Pearson; nabeel.qazi@pearson.com
Qi, Yi; Educational Testing Service; yqi@ets.org
Qiao, Xin; Southern Methodist University; xqiao@smu.edu
Qin, Qi; Gwinnett County Public Schools; qinqi715@gmail.com
Qu, Yanxuan; ETS; yqu@ets.org
Quan, Jia; ; jia.quan@ufl.edu
Quan, Yale; University of Washington; yalequan@uw.edu
Quanbeck, Mari; University of Minnesota - Twin Cities; quanb016@umn.edu
Quansah, Frank; University of Cape Coast; frank.quansah1@stu.ucc.edu.gh
Quesen, Sarah; WestEd; sarah.quesen@gmail.com
Quinones Perez, Isaris; University of Puerto Rico- Rio Piedras; isaris.quinones@upr.edu
Quirk, Victoria L.; University of Illinois at Urbana-Champaign; vquirk3@illinois.edu
Rabe-Hesketh, Sophia; University of California, Berk; sophiarh@berkeley.edu
Rabinowitz, Stanley N; Edmetric LLC; snr55aus@gmail.com
Rafferty, Anna; ; arafferty@carleton.edu
Rahal, Charles; University of Oxford; charles.rahal@sociology.ox.ac.uk
Ramsawak-Jodha, Nalini; University of The West Indies, St. Augustine, Trinidad; nalini.ramsawak-jodha@sta.uwi.edu
Randall, Jennifer; University of Massachusetts; jrandall@educ.umass.edu
Rao, Analía; University of California-Irvine; aerao@uci.edu
Ravelo, Guillermo; Boston College; ravelog@bc.edu
Raymond, Krystina; University of Toronto; krystina.raymond@utoronto.ca
Reckase, Mark; Psychometric Solutions; reckase@msu.edu
Redman, Elizabeth; UCLA CRESST; redman@cresst.org
Reichert, Frank; The University of Hong Kong; reichert@hku.hk
Reilly, Amy; Pearson; amy.reilly@pearson.com
Ren, Hao; Pearson; h.ren@hotmail.com
Ren, Jinglei; University of Maryland; jinglei@umd.edu
Rhemtulla, Mijke; University of California, Davis; mrhemtulla@ucdavis.edu
Rijmen, Frank; Cambium Assessment, Inc; frank.rijmen@cambiumassessment.com
Rikoon, Sam; ; srikoon@ets.org
Rios, Joseph A.; University of Minnesota; jrios@umn.edu
Robar, Julianne; Renaissance Learning; julianne.robar@renaissance.com
Robb, Colleen; Altus Assessments; crobb@altusassessments.com
Roberts, Jeremy; PBS KIDS Digital; jdroberts@pbs.org
Robin, Frederic; ETS; frobin@ets.org
Rodriguez, Karina; Highlander Institute; krodriguez@highlanderinstitute.org
Rodriguez, Michael C.; University of Minnesota; mcrdz@umn.edu
Roebeck, Mark; Pearson; mark.Roebeck@pearson.com
Roeber, Edward Dean; Michigan Assessment Consortium; roeber@msu.edu
Rollins, Taj; ; Tarol3@morgan.edu
Rome, Logan; Curriculum Associates; lrome@cainc.com
Rosales de Véliz, Leslie Vanessa; Jml Measurement and Testing Services, LLC; leslie.rosales@gmail.com
Rosca, Oxana; University at Albany - SUNY; orosca@albany.edu
Rosen, Yigal; BrainPop; yigalr@brainpop.com
Rosenberg, Sharyn; NAGB; sharyn.rosenberg@ed.gov
Ross, Linette P.; NBME; lross@nbme.org
Rotou, Ourania; New Meridian;
Roussos, Louis; Cognia; louis.roussos@cognia.org
Rozunick, Chris; TEA; christine.rozunick@tea.texas.gov
Ruan, Chunyi; ETS; CRuan@ets.org
Rubright, Jonathan; National Board of Medical Examiners; jrubright@nbme.org
Ruiz-Primo, Maria Araceli; Stanford University; aruiz@stanford.edu

Participant Emails

(Last Name, First Name; Affiliation; Email)

Runge, Andrew; Duolingo; andrew_runge@duolingo.com
Runyon, Christopher; NBME; CRunyon@nbme.org
Russell, Michael; Boston College; michael.russell@bc.edu
Russell, Michael; Boston College; russelmh@bc.edu
Rust, Keith; Westat; rustk1@westat.com
Rutkowski, David; Indiana University; rustk1@westat.com
Rutkowski, Leslie; Indiana University; lrutkows@iu.edu
Ryu, Ehri; Boston College; ehri.ryu@bc.edu
Sabatini, John; Institute for Intelligent Systems, University of Memphis; jpsbtini@memphis.edu
Saeidzadeh, Seyedehtanaz; University of Alberta; saeidzad@ualberta.ca
Sahin, Fusun; American Institutes for Research; fsahin@air.org
Salas, Jorge; Vanderbilt University; jorge.a.salas@vanderbilt.edu
Sales, Adam C; Worcester Polytechnic Institute; asales@wpi.edu
Sambucharan-Mohammed, Murella; The University of The West Indies; murella.sambucharan-mohammed@gmail.com
Sammit, George; Southern Methodist University; gsammit@smu.edu
Sanchez, Edgar; ACT; edgar.sanchez@act.org
Sanchez, Edgar; ACT; Suppression305@gmail.com
Sanders, Elizabeth A.; University of Washington, Seattle; lizz@uw.edu
Sarac, Merve; UW-Madison; sarac@wisc.edu
Sarraf, Shimon; Indiana University; ssarraf@indiana.edu
Satkus, Paulius; Graduate Management Admission Council; psatkus@gmac.com
Sato, Edynn; Sato Education Consulting LLC; edynn@satoeducationconsulting.com
Sato, Yoshikazu; Admission Center, Kyushu University; ysato@artsci.kyushu-u.ac.jp
Sayin, Ayfer; ; sayinayfer@gmail.com
Schaefer, Katarina; James Madison University; schae2ke@jmu.edu
Schaller, Nils-Jonathan; IPN – Leibniz Institute for Science and Mathematics Education; schaller@leibniz-ipn.de
Schellman, Madeline; ; mas13@uga.edu
Schmidt, Amy; Pearson; aschmidt19@comcast.net
Schmidt, Amy Elizabeth; Pearson; amy.schmidt@pearson.com
Schneider, Christina; Cambium Assessment, Inc.; christina.schneider@cambiumassessment.com
Schneider, Wei S.; The College Board; wschneider@collegeboard.org
Schöber, Christian; IfBQ Hamburg; Christian.Schoeber@iqsh.landsh.de
Schoemann, Alexander M; East Carolina University; schoemanna@ecu.edu
Schonberg, Christina; IXL Learning; cschonberg@ixl.com
Schor, David; ETS; dschor@ets.org
Schultz, Matthew; AICPA; matthew.schultz01@gmail.com
Schumacker, Randall; ; rschumacker@ua.edu
Schwartz, Robert; ACT; bob.schwartz@act.org
Sgammato, Adrienne; ETS; asgammato@ets.org
Shafer Willner, Lynn; University of Wisconsin - Madison; bearwolf88@gmail.com
Shapovalov, Yelisey A; ; shapovyx@dukes.jmu.edu
Shapovalov, Yelisey A.; James Madison University; shapovyx@jmu.edu
Sharairi, Sid; Riverside Insights; sid.sharairi@riversideinsights.com
Shaw, Emily; College Board; eshaw@collegeboard.org
Shear, Benjamin R.; University of Colorado Boulder; benjamin.shear@colorado.edu
Shen, Yawei; ; yawei.shen@pearson.com
Sheng, Yanyan; University of Chicago; y.sheng@uchicago.edu
Shepard, Lorrie Ann; University of Colorado Boulder; Lorrie.Shepard@Colorado.edu
Shermis, Mark David; Performance Assessment Analytics, LLC; mshermis@gmail.com
Shetty, Sandeep; AIR; sshetty@impaqint.com
Shi, Dexin; University of South Carolina; shid@mailbox.sc.edu
Shi, Qingzhou; University of Alabama; qshi7@crimson.ua.edu
Shibayama, Tadashi; Tohoku University; sibayama@tohoku.ac.jp
Shih, Ching-Lin; National Sun Yat-Sen University; educls@g-mail.nsysu.edu.tw
Shin, David; Pearson; cshin0803@gmail.com
Shin, David; Pearson; david.shin@pearson.com
Shin, Hyo Jeong; Educational Testing Service; hshin@ets.org
Shin, Jinnie; ; jinnie.shin@coe.ufl.edu
Shin, Minkyoung; Seoul Women's University; sweet068@hanmail.net
Shin, Nami; ATLAS, University of Kansas; namishin@ucla.edu
Shivraj, Pooja; American Board of OBGYN; pshivraj@abog.org
Shono, Yusuke; Claremont Graduate University; yusuke.shono2@cgu.edu
Shrestha, Shovana; University of Alberta; shovana@ualberta.ca
Sikali, Emmanuel; ; Emmanuel.Sikali@ed.gov
Sinharay, Sandip; Educational Testing Service; ssinharay@ets.org
Sinval, Jorge; Iscte — University Institute of Lisbon; jorge.sinval@iscte-iul.pt
Sireci, Stephen; University of Massachusetts Amherst; sireci@acad.umass.edu

Participant Emails

(Last Name, First Name; Affiliation; Email)

Sireci, Stephen G; University of Massachusetts, Amherst; sireci@umass.edu
Sitarenios, Gill; Altus Assessments; gsitarenios@altusassessments.com
Skoblow, Hanamori; University of Missouri; skoblowh@mail.missouri.edu
Skorupski, William; Data Recognition Corp; william_skorupski@yahoo.com
Slater, Sharon; ETS; sslater@ets.org
Smith, Larissa; NBOME; LSmith@nbome.org
Smith, Thomas M.; Vanderbilt University; thomas.smith.1@vanderbilt.edu
Snow, Erica; Roblox Corp; esnow@roblox.com
Soland, James; University of Virginia; jgs8e@virginia.edu
Solano-Flores, Guillermo; Stanford University; gsolanof@stanford.edu
Someshwar, Shonai; UNC Greensboro; s_somesh@uncg.edu
Somsong, Sarunya; Srinakharinwirot University & Southern Methodist University; ssomsong@smu.edu
Song, Jieqing; Foreign Language Teaching and Research Press; jsong@fltrp.com
Song, Yi; Educational Testing Service; ysong@ets.org
Song, Yoon Ah; Center for Applied Linguistics; episteme84@hotmail.com
Song, Yuting; University of Alberta; yuting.song@ualberta.ca
Sonnleitner, Philipp; Luxembourg Centre for Educatio; philipp.sonnleitner@uni.lu
Sorrel, Miguel A.; Universidad Autónoma de Madrid; miguel.sorrel@uam.es
Sotelo, Jose de Jesus; Educational Testing Service (ETS); josesotelo2023@u.northwestern.edu
Springer, Matthew; UNC; mgspringer@unc.edu
Stark, Stephen; University of South Florida; sestark@usf.edu
Steedle, Jeffrey; ACT, Inc.; jtsteedle@gmail.com
Steimel, Kenneth; ETS; ksteimel@ets.org
Steinberg, Jonathan; Test; jsteinberg@ets.org
Stewart, John; Amplify Education; jds@amplify.com
Stoeger, Jordan Nelson; Data Recognition Corporation; jstoeger@datarecognitioncorp.com
Stopek, Joshua; AICPA; Joshua.stopek@aicpa-cima.com
Stout, William; University of Illinois; w-stout1@illinois.edu
Struthers, Vince; Data Recognition Corporation; VStruthers@datarecognitioncorp.com
Student, Sanford; University of Colorado Boulder; sanford.student@colorado.edu
Su, Shiyang; University of Central Florida; Shiyang.Su@ucf.edu
Su, Yu-Lan; Ascend Learning; suyulan@gmail.com
Suárez-Álvarez, Javier; University of Massachusetts Amherst; suarezj@umass.edu
Subedi, Dipendra; Pearson; dipendra.subedi@pearson.com
Subedi, Dipendra; Pearson Assessments; dipendrasubedi@gmail.com
Subhiyah, Raja G; National Board of Medical Examiners; rsubhiyah@nbme.org
Suh, Hongwook; Cambium Assessment, Inc.; hongwooks@gmail.com
Suh, Yon Soo; NWEA; yon.soo.suh@nwea.org
Suh, Yon Soo; UCLA; ysuh09@ucla.edu
Suk, Youmi; Teachers College Columbia University; ysuk@tc.columbia.edu
Sum, Jane; A*STAR; jane_sum@imcb.a-star.edu.sg
Sun, Sundance Zhihong; The University of Melbourne; zhihongs@student.unimelb.edu.au
Sussman, Joshua; UC Berkeley; jsussman@berkeley.edu
Svensson, Cicek; Caveon Test Security; cicek.svensson@caveon.com
Svetina Valdivia, Dubravka; Indiana University; dsvetina@indiana.edu
Swain, Matthew; American Board of Internal Medicine; mswain@abim.org
Swaminathan, Hariharan; University of Connecticut; swami@uconn.edu
Sweet, Tracy; University of Maryland College Park; tsweet@umd.edu
Swygert, Kimberly; National Board of Medical Examiners; kswygert@nbme.org
Talreja, Vinita; AWS; vgtalrej@amazon.com
Tan, Bin; University of Alberta; btan4@ualberta.ca
Tan, Sinon; Duolingo; sinon@duolingo.com
Tan, Xuan (Adele); ETS; atan@ets.org
Tan, Yanyan; ; yt58945@uga.edu
Tang, Nai-En; ; naientang@gmail.com
Tang, Steven; eMetric; steven@emetric.net
Tang, Xiuxiu; ; tang469@purdue.edu
Tao, Shuqin; Cambium Assessment, Inc.; shuqin.tao@gmail.com
Taylor, Catherine; ; ctaylor@uw.edu
Teitelbaum, Jeremy; University of Connecticut; Jeremy.Teitelbaum@uconn.edu
Telfort, Roseline; ; Rtelfo2@lsu.edu
Templin, Jonathan; University of Iowa; jonathan-templin@uiowa.edu
Thiessen, Brad; ; bradthiessen@mac.com
Thomas, Elizabeth R.; Southern Methodist University; thomaser@smu.edu
Thomas, Jay; ACT, Inc; jay.thomas@act.org
Thompson, Kathryn Nicole; James Madison University; thompskn@jmu.edu
Thompson, W. Jake; University of Kansas; wjakethompson@gmail.com

Participant Emails

(Last Name, First Name; Affiliation; Email)

Thompson, William Jacob; University of Kansas; jakethompson@ku.edu
Thum, Yeow Meng; NWEA; yeow.meng@nwea.org
Thurlow, Martha; National Center On Educational; thurl001@umn.edu
Tian, Chen; ; 736218349@qq.com
Tijmstra, Jesper; Tilburg University; j.tijmstra@uvt.nl
Tingir, Seyfullah; Amplify Education; stingir@amplify.com
Todd, Jessica Andrews; Educational Testing Service; jandrewstodd@ets.org
Tolentino, Lissette; ; ltolen@ufl.edu
Tomkowicz, Joanna; Data Recognition Corporation; jtomkowicz@datarecognitioncorp.com
Tong, Ye; National Board of Medical Examiners; yetong@nbme.org
Topczewski, Anna; Law School Admission Council; atopczewski@lsac.org
Topczewski, Anna; Law School Admission Council; topczewski.anna@gmail.com
Toptas, Cigdem ; ciggdemtoptas@gmail.com
Torres Irribarra, David; Pontificia Universidad Católica De Chile; davidtorres@uc.cl
Torres, Chloe; NWEA; chloe.torres@nwea.org
Toton, Sarah Linnea; Caveon Test Security; sarah.toton@caveon.com
Toyama, Yukie; University of California, Berkeley; yukie.toyama@berkeley.edu
Traynor, Anne; Purdue University; atraynor@purdue.edu
Tsai, Chia-Lin; ; chialin.tsai@unco.edu
Tu, Naidan; University of South Florida; naidantu@usf.edu
Tubbs Dolan, Carly; New York University; carly.tubbs@nyu.edu
Tufail, Mubeshera; ; mubesheratufail@yahoo.com
Turner, Ronna; University of Arkansas; rcturner@uark.edu
Twing, Jon S; Pearson; jon.twing@pearson.com
Tyack, Lillian; Boston College; tyack@bc.edu
Ulitzsch, Esther; Leibniz Institute for Science and Mathematics Education; ulitzsch@inp.uni-kiel.de
Underwood, Sarah; Florida Department of Education; Sarah.Underwood@fldoe.org
Valdivia Medinaceli, Montserrat B; Indiana University Bloomington; mbvaldiv@iu.edu
van der Linden, Wim J; University of Twente; wjvdlinden@outlook.com
van Rijn, Peter; ETS Global; pvanrijn@etsglobal.org
Vanacore, Kirk P; Worcester Polytechnic Institute; kpvanacore@wpi.edu
Vasquez-Colina, Maria; Florida Atlantic University; mvasque3@fau.edu
Veldkamp, Bernard; University of Twente; b.p.veldkamp@gw.utwente.nl
Verges, Vince; Florida Department of Education; vergesvincent@gmail.com
Vernon, Annette; University of Iowa; annette-vernon@uiowa.edu
Vispoel, Walter; University of Iowa; walter-vispoel@uiowa.edu
Vo, Thao; Washington State University; thao.vo@wsu.edu
Vo, Yen; University of Iowa; yen-vo@uiowa.edu
von Davier, Alina A; Duolingo; avondavier@duolingo.com
von Davier, Matthias; Boston College; Matthias.vonDavier@bc.edu
Von Davier, Matthias; Boston College; vondavim@bc.edu
Vorenkamp, Ellen; Michigan Assessment Consortium; vorenke83@gmail.com
Walker, Marcus; National Commission On Certification of Physician Assistants; marcusw@nccpa.net
Walker, Michael E.; Educational Testing Service; mwalker@ets.org
Walkowiak, Templ A.; NC State University; tawalkow@ncsu.edu
Wall, Nathan; eMetric; nwall@emetric.net
Wallace, Leslie; Westat; LeslieWallace@westat.com
Wallin, Gabriel; ; g.a.wallin@lse.ac.uk
Walsh, Cole; Altus Assessments; cwalsh@altusassessments.com
Walters, Kristy; Corunna Public Schools; kwalters@corunna.k12.mi.us
Wan, Ping; Data Recognition Corporation; PWan@datarecognitioncorp.com
Wan, Siyu; ABIM; siyuwan93@outlook.com
Wang, Aijun; FSBPT; wajlm2003@gmail.com
Wang, Bowen; University of Florida; bowen.wang@ufl.edu
Wang, Changjiang; Pearson; Changjiang.Wang@Pearson.com
Wang, Chuang; University of Macau; wangc@um.edu.mo
Wang, Chun; University of Washington; wang4066@uw.edu
Wang, Chunxin; ; annwang728@gmail.com
Wang, Dongran; Tobii Electronic Technology Suzhou Co., Ltd; Murphy.wang@tobii.com
Wang, Dongwei; UMass Amherst; dongweiwang@umass.edu
Wang, Huan; Data Recognition Corporation; HWang@DataRecognitionCorp.com
Wang, Jincai; Suzhou University; wwwanggj@163.com
Wang, Kuo; Southern Methodist University; wangp@mail.smu.edu
Wang, Kuo; Southern Methodist University; wangp@smu.edu
Wang, Nixi; University of Washington; nixiwang@uw.edu
Wang, Selena; Yale School of Public Health; selenashuowang@gmail.com
Wang, Serena; UC, Berkeley; serenalwang@berkeley.edu
Wang, Shichao; ACT; shichao.wang@act.org

Participant Emails

(Last Name, First Name; Affiliation; Email)

Wang, Shiyu; ; swang44@uga.edu
Wang, Shudong; NWEA; shudong.wang@NWEA.org
Wang, Songtao; Oise/University of Toronto; songtaowang@uvic.ca
Wang, Ting; American Board of Family Medicine; twang@theabfm.org
Wang, Xi; ; xwang405@uiowa.edu
Wang, Xiaolin; Pearson VUE; xwangiub@gmail.com
Wang, Yihao; Southern Methodist University; yihaow@mail.smu.edu
Wang, Yu; University of Minnesota, Twin Cities; wang7919@umn.edu
Wang, Yuan; ETS; ywang@ets.org
Wang, Yurou; University of Alabama; ywang329@ua.edu
Wang, Zhuoran; National Council of State Boards of Nursing (NCSBN); wzhranran@gmail.com
Warner, Zachary; New York State Education Dept.; zacharybwarner@yahoo.com
Way, Denny; College Board; dway@collegeboard.org
Webb, Blue; American Institutes for Research; bwebb@air.org
Weber, Anjali; Amazon Web Services (AWS); anjaliw@amazon.com
Weeks, Jonathan; Educational Testing Service; jweeks@ets.org
Weese, James; University of Arkansas; jweese@uark.edu
Wei, Xin; SRI International; xin.wei@sri.com
Wei, Youhua; Educational Testing Service; ywei@ets.org
Weir, John; National Commission On Certification of Physician Assistants; johnw@nccpa.net
Weiss, David; University of Minnesota; djweiss@umn.edu
Weissman, Alexander; Law School Admission Council; aweissman@lsac.org
Welch, Catherine; University of Iowa; catherine-welch@uiowa.edu
Wellberg, Sarah; University of Colorado, Boulder; Sarah.Wellberg@colorado.edu
Wells, Craig; UMass Amherst; cswells@umass.edu
Welsh, Megan; University of California, Davi; welsh.megan@gmail.com
Wheeler, Jordan M.; University of Georgia; jmwheeler@uga.edu
Wheeler, Kelley; ACS Ventures, LLC; kelleyrwheeler@gmail.com
Whitmer, John; Institute of Education Statistics; john.whitmer@ed.gov
Whittaker, Tiffany; University of Texas; t.whittaker@austin.utexas.edu
Wiberg, Marie; Umeå University; marie.wiberg@umu.se
Wibowo, Arianto; Measurement Incorporated; awibowo@measinc.com
Wieman, Carl; Stanford University; cwieman@stanford.edu
Wiley, Andrew; ACS Ventures; awiley@acsventures.com
Wilhelm, Anne G.; Southern Methodist University; awilhelm@mail.smu.edu
Williams, Kevin; Educational Testing Service; kmwilliams@ets.org
Wilson, Jonee; NC State University; jwilson9@ncsu.edu
Wilson, Joshua; University of Delaware; joshwils@udel.edu
Wilson, Mark; Berkeley School of Education, UC Berkeley; markw@berkeley.edu
Wind, Stefanie A.; University of Alabama; swind@ua.edu
Wine, Marjorie; ATLAS: University of Kansas; mwine122@gmail.com
Winter, Sonja D; University of Missouri; sdwinter@missouri.edu
Winters, Erin; ; eewinters@ucdavis.edu
Wise, Steven; NWEA; steve.wise@nwea.org
Witmer, Sara; Michigan State University; switmer@msu.edu
Wolf, Mikyung Kim; Educational Testing Service; mkwolf@ets.org
Wollack, James; University of Wisconsin; jwollack@wisc.edu
Wong, Yun Leng; University of Minnesota; wong0620@umn.edu
Wongvorachan, Tarid; University of Alberta; wongvora@ualberta.ca
Woo, Ada; Ascend Learning; adawoo811@gmail.com
Woo, Yejin; Ewha Womans University; wooye6093@ewhain.net
Woods, Carol; Cambium Assessment; carol.woods@cambiumassessment.com
Workman, Trent; Pearson; Trent.Workman@Pearson.com
Wraga, Bill; University of Georgia; wraga@uga.edu
Wu, Amery; University of British Columbia; amery.wu@ubc.ca
Wu, Yi-Chen; University of Minnesota/NCEO; wuxx0207@umn.edu
Wu, Yi-Fang; Cambium Assessment, Inc.; wuyifang91@gmail.com
Wyse, Adam E; Adam Wyse; adam.wyse@renaissance.com
Xi, Nuo; Educational Testing Service; nxi@ets.org
Xiao, Houping; Georgia State University; hxiao@gsu.edu
Xiao, Jiaying; University of Washington; jxiao6@uw.edu
Xiao, Xingyao; XXY; xiaoxg@berkeley.edu
Xie, Aolin; Law School Admission Council; olymxie@gmail.com
Xie, Benjamin; University of Washington; bxie@uw.edu
Xin, Tao; Beijing Normal University; xtao_bnu@163.com
Xiong, Jiawei; Pearson; jiawei.xiong@pearson.com
Xiong, Yao; Roblox Corporation; xy.xiongyao@gmail.com

Participant Emails

(Last Name, First Name; Affiliation; Email)

Xu, Gongjun; University of Michigan; gongjun@umich.edu
Xu, Guanlan; ; guanlan-xu@uiowa.edu
Xu, Lingling; Peking University; xllpsy@qq.com
Xu, Menglin; ; xumenglin920@gmail.com
Xu, Rujun; University of Virginia; bwp7ab@virginia.edu
Xu, Xiangyu; University of Notre Dame; xxu11@nd.edu
Xu, Xiaochen; University of California, Davis; xixu@ucdavis.edu
Yablonski, Maya; Stanford University; mayay@stanford.edu
Yan, Duanli; ETS; dyan@ets.org
Yan, Wei Wei; Altus Assessments; wyan@altusassessments.com
Yan, Yan; Georgia Tech; yany@gatech.edu
Yancey, Kevin; Duolingo; kyancey@duolingo.com
Yang, Hongyu; ; hyang3@umd.edu
Yang, Ji Seung; University of Maryland; jsyang@umd.edu
Yang, Zhitong; ; zyang@ets.org
Yao, Lihua; Northwestern University; yaolihua081@gmail.com
Yavuz Temel, Güler; Hamburg University; gueler.yavuz.temel@uni-hamburg.de
Yavuz, Sinan; American Institutes for Research; syavuz@air.org
Ye, Daisy; DRC; dye@datarecognitioncorp.com
Yeatman, Jason; Stanford University; jyeatman@stanford.edu
Yen, Shu Jing; Center for Applied Linguistics; syen@cal.org
Yeum, Sung Kun; Geunhwa Girls High School; ballum11@naver.com
Yigit, Hulya Duygu; ; yigit.hulyad@gmail.com
Yildirim-Erbasli, Seyma N.; Concordia University of Edmonton; seyma.yildirim-erbasli@concordia.ab.ca
Ying, Zhiliang; Columbia University; zying@stat.columbia.edu
Yoo, Hanwook; Educational Testing Service; hanuki82@gmail.com
Young, Mackenzie; Cambium Assessment; mackenzie.young@cambiumassessment.com
Younger, Jessica; PBS KIDS; jryounger@pbs.org
Yuan, Ye; University of Georgia; yeyuan0106@gmail.com
Yumsek-Akbaba, Meltem; Ministry of Education, Turkey; myumsek@gmail.com
Zahner, Doris; CAE; dzahner@cae.org
Zan, Yixin; George Mason University; yzan2@gmu.edu
Zapata-Rivera, Diego; Educational Testing Service; dzapata@ets.org
Zavitkovsky, Paul; Center for Urban Education Leadership, University of Illinois at Chicago; pzavit@uic.edu
Zeng, Lin; Louisiana State University; zlin8@lsu.edu
Zenisky, April; University of Massachusetts Amherst; azenisky@educ.umass.edu
Zhang, Fan; University of Delaware; fzhang@udel.edu
Zhang, Jihong; University of Iowa; jihong-zhang@uiowa.edu
Zhang, Jinming; University of Illinois at Urba; jmzhang@illinois.edu
Zhang, Lijin; Stanford University; lijinzhang@stanford.edu
Zhang, Litong; DRC; LZhang@DataRecognitionCorp.com
Zhang, Mengyao; National Conference of Bar Examiners; mzhang@ncbex.org
Zhang, Mingqin; University of Iowa; mingqin-zhang@uiowa.edu
Zhang, Mo; Educational Testing Service; mzhang@ets.org
Zhang, Oliver; College Board; ozhang@collegeboard.org
Zhang, Susu; University of Illinois at Urbana-Champaign; szhan105@illinois.edu
Zhang, Ting; American Institutes for Research; tzhang@air.org
Zhang, Tongxin; ; ztx199828@163.com
Zhang, Wenqing; East China Normal University (Intern, Beijing Insight Online Management Consulting Co.,Ltd); zhang.psychometrics@foxmail.com
Zhang, Xiao; DRC; xzhang@datarecognitioncorp.com
Zhang, Yichi; ; yzhang97@usc.edu
Zhang, Yifan; ; zyf2020@connect.hku.hk
Zhang, Yu; Federation of State Boards of Physical TherapEy; yzhang@fsbpt.org
Zhang, Zhonghua; University of Melbourne; zhonghua.zhang@unimelb.edu.au
Zhao, Mingye; National Board of Osteopathic Medical Examiners; mzhao@nbome.org
Zhao, Xinchu; Roblox Corporation; xinchuz@gmail.com
Zhao, Yang; American Board of Internal Med; yzhao@abim.org
Zhao, Yu (Tracy); ; tracy.zhao@pearson.com
Zheng, Bin; University of Alberta; bzheng1@ualberta.ca
Zheng, Chanjin; East China Normal University; chjzheng@dep.ecnu.edu.cn
Zheng, Guoguo; Amplify Education; gzheng@amplify.com
Zheng, Xiaying; American Institutes for Research; xzheng@air.org
Zheng, Yi; Arizona State University; yi.isabel.zheng@asu.edu
Zhong, Xiaoting; University of Iowa; xiaoting-zhong@uiowa.edu
Zhou, Junfan; Hong Kong Polytechnic University; junfan.zhou@connect.polyu.hk
Zhou, Todd; Univeristy of Maryland; toddzhou7@gmail.com

Participant Emails

(Last Name, First Name; Affiliation; Email)

Zhou, Xuechun; Ascend Learning; xuechun.zhou@ascendlearning.com
Zhu, Danqi; ; danqi.zhu@bc.edu
Zhu, Danqi; Fordham University; dzhu17@fordham.edu
Zhu, Ruoyi; University of Washington; zhux0445@uw.edu
Ziedzor, Reginald; ; rziedzor@amplify.com
Ziedzor, Reginald; University of Southern Illinois Carbondale; ziedzor@siu.edu
Ziegler, Garrett; College Board; gziegler@collegeboard.org
Zilberberg, Anna; ; azilberb@gmail.com
Zor, Selay; University of Georgia; selayzor@gmail.com
Zor, Selay; University of Georgia; sz37952@uga.edu
Zou, Zhimin; Wenzhou University; zhiminzou@sina.com
Zu, Jiyun; Educational Testing Service; jzu@ets.org
Zumbo, Bruno D.; University of British Columbia; bruno.zumbo@ubc.ca

National Council on Measurement in Education is very grateful to the following organizations for their generous financial support of our 2023 Annual Meeting.

GOLD

CollegeBoard

duolingo
english test

ETS®

edCount LLC®
because all students count

Pearson

WALTON FAMILY
FOUNDATION

SILVER

ACS VENTURES
BRIDGING THEORY & PRACTICE

Alpine
Testing Solutions

AAMC

Center for Assessment

cognia®

COLLEGE OF EDUCATION
+ HUMAN DEVELOPMENT
UNIVERSITY OF MINNESOTA

Curriculum Associates®

New Meridian

NBME®

nwea® believe in what's possible®

FRIENDS

BEAR Center
Berkeley Evaluation & Assessment Research Center

HARVARD
GRADUATE SCHOOL OF EDUCATION

UCLA | CRESST

UMassAmherst
College of Education
Center for Educational Assessment

OTHER SUPPORTERS

BUROS
CENTER FOR TESTING

flexMIRT®

Graduate Management Admission Council™

IOWA™

edCount is pleased to sponsor NCME

Celebrating 20 years of dedication to equity and excellence for all students

Our Belief Statement

At edCount, we believe that All Students Count.

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

- We understand the immense opportunity we have as innovators in the educational assessment field and acknowledge that historical approaches to measurement have contributed to deeply ingrained systems of inequity for historically marginalized students.
- We acknowledge that one's environment directly impacts the quality of one's work.
- We are committed to building and fostering a workforce that reflects the diverse populations that we serve.

Our Services

Assessment Implementation and Evaluation

- Assessment System Design and Evaluation
- Score Reporting and Performance Standards
- Policies and System Documentation for Federal Peer Review
- Alignment Evaluations, Validity Evaluations, and Cognitive Labs
- Performance Level Descriptor Studies, Instructional Use Studies, and Learner Characteristics Studies

Instructional Systems and Capacity Building

- Curriculum Development
- Standards Implementation
- Standards, Curriculum, Instruction, and Culture Studies
- Capacity Building for Educators
- Evaluating Instructional Systems

Policy Analysis and Technical Assistance

- Federal Policy Evaluation
- Application Review for Federal Competitions and Programs
- Management of Multi-State Policy Response Collaboratives
- Expert Policy Analysis and Guidance
- Technical Advisory Committees (TACs)

edCount LLC®
because all students count

1800 Diagonal Road, Suite 600 | Alexandria, VA 22314
www.edCount.com | (202) 895-1502 | info@edCount.com