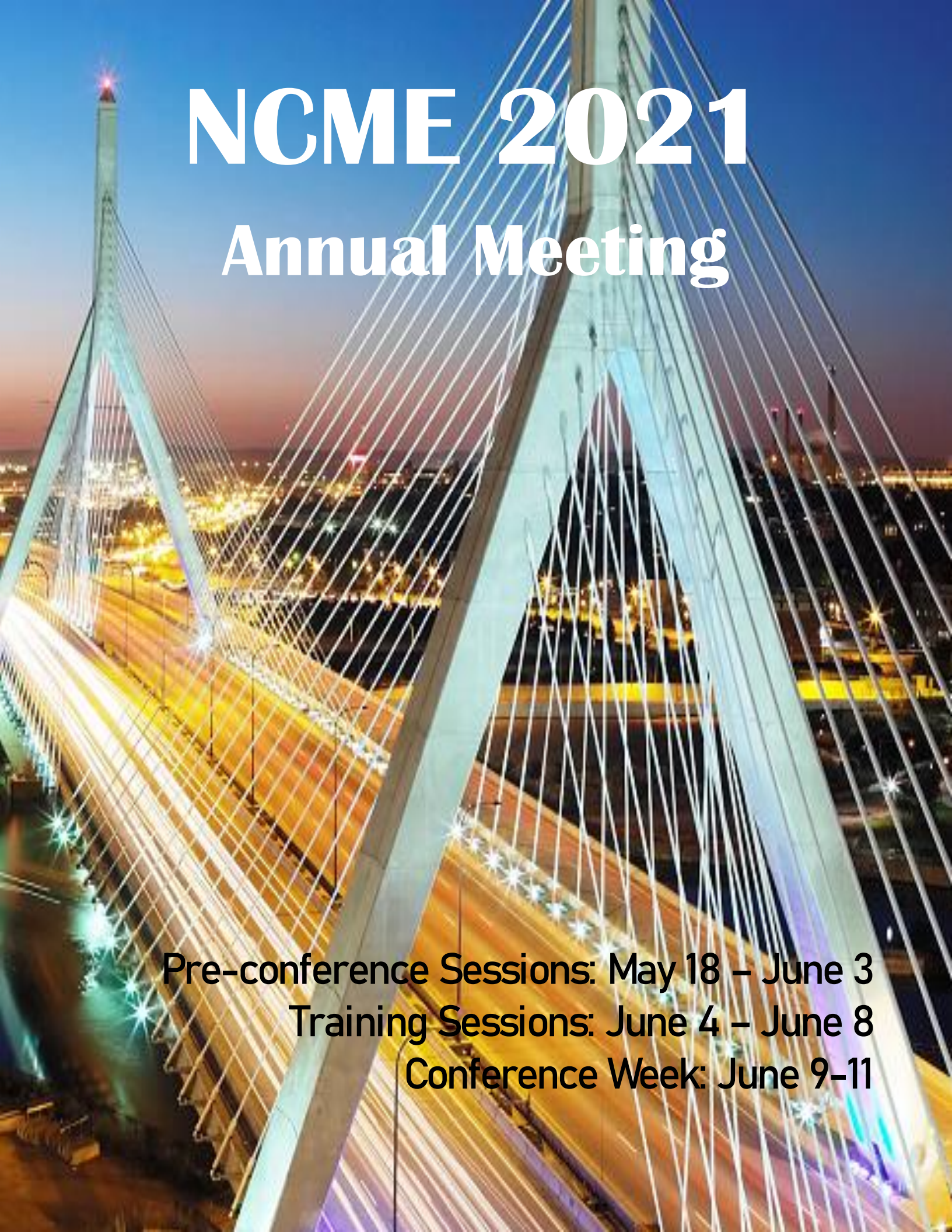

NCME 2021

Annual Meeting

Pre-conference Sessions: May 18 – June 3
Training Sessions: June 4 – June 8
Conference Week: June 9-11

Welcome from the Program Chairs

Welcome to the 2021 NCME Annual Meeting! The past year has been challenging for everyone. We have learned to be flexible and adapt to the ever-changing circumstances that have occurred on a monthly, weekly, and sometimes daily basis. This experience was no different when it came to the planning of this year's meeting and program. Most of us have grown accustomed to the virtual meeting format – raise your hand if you had not been in a Zoom breakout room, played with virtual backgrounds, or used video filters ("I'm not a cat!") before last year! While we acknowledge that virtual interactions cannot replace the authentic experience of face-to-face conversations and discussions, we have structured the program to take advantage of the benefits and conveniences of the virtual format. For example, we are offering several pre-conference sessions highlighting some amazing content. The pre-conference sessions begin during the week of May 18, every Tuesday and Thursday, leading up to the week of the conference, June 8 to 11.

This year's conference theme is "Bridging Research and Practice". Research is the foundation of our field and continues to shape and advance our industry. Practice is where we can reach millions of people, and where critical decisions, such as certification, admissions, and placements, are made. Our goal is to encourage and foster a stronger bond between research and practice. Research needs to help address practical challenges and implications; practice needs to be grounded in research. To support this theme, we planned a series of invited theme sessions including:

- (Past) *Lessons about the modeling and measurement of human abilities*, Tuesday, May 18, 11:00am-12:30pm ET,
- (Present) *Stakeholder perspectives on validating licensure examinations*, Thursday, June 10, 1:00pm-2:00pm ET,
- (Present) *Assessments for different purposes: issues on scoring, score use, and measurement*, Thursday, June 10, 2:15pm-3:45pm ET,
- (Present) *Pivoting in a pandemic*, Thursday, June 10, 4:00pm-5:30pm ET,
- (Future) *Where Do We Go from Here? A practitioner's discussion of our post-pandemic world*, Friday, June 11, 1:00pm-2:00pm ET,
- (Future) *Looking ahead – Bridging future research and practice in credentialing*, Friday, June 11, 1:00pm-2:00pm ET.

In addition, we have several "hot topics" invited sessions:

- *Using longitudinal assessment to support professional development*, Thursday, May 20, 11:00am-12:30pm ET,
- *Education literacy for psychometricians*, Thursday, May 27, 2:00pm-3:30pm ET,
- *Artificial Intelligence, Machine Learning, and the Future of Assessment*, Wednesday, June 9, 9:00am-10:30am ET,
- *The value of assessment data from spring 2021: A debate*, Wednesday, June 9, 1:00pm-2:00pm ET
- *Lessons learned from the pandemic: how do credentialing programs prepare for the next major crisis/disruption?* Friday, June 11, 9:00am-10:30am ET
- *Remembering "career" in college and career readiness*, Friday, June 11, 9:00am-10:30am ET
- *The future of college admissions testing*, Friday, June 11, 2:15pm-3:45pm ET

Another meeting highlight is the featured session for NCME's Committee on Diversity in Testing, *Black Lives Matter in Educational Measurement* (Wednesday, June 9, 11:15am-12:45pm). We also have a session devoted to discussing the barrier and opportunities for women in the measurement field, *Advancing Women in Measurement: Barriers and Opportunities* (Tuesday, May 25, 2:00pm-3:30pm). In addition, we have a session dedicated to Edmund Gordon, *Using Educational Assessments to Educate: Opportunities for Leveraging the "Power" of Assessment* (Wednesday, June 9, 9:00am-10:30am), and a session organized by the National Association of Assessment Director (NAAD), *Assessment Literacy: Practical Applications and Implications* (Wednesday, June 9, 2:15pm-3:45pm). Also, be sure to check out the sessions organized by the various NCME SIGIMIEs (Special Interest Groups in Measurement in Education – marked with 'SIGIMIE' in the title). Lastly, even the pandemic will not stop us from continuing several NCME Annual Meeting traditions such as the *NCME Business Meeting and Presidential Address* on Thursday, June 10, 10:45am-12:45pm, the *NCME Fitness Run*, as well as yoga and meditation sessions.

We must acknowledge the incredible and talented cadre of NCME members and colleagues who generously volunteered their time and expertise this year to ensure that we will have a high-quality program. Many of you reviewed proposals and provided very helpful feedback; and many of you volunteered and are serving as chairs and discussants for the program. We cannot thank you enough! We are also grateful to Ye Tong (NCME President) and the NCME Board; Sarah Quesen

(Training & Professional Development Committee Chair) and her committee; Maura O'Riordan and Scott Holcomb (the Graduate Student Committee Co-chairs); and Erin O'Leary and Ethan Gray (Talley, NCME's management partner).

Last but certainly not least, we would like to acknowledge the 2020 NCME Annual Meeting Co-Chairs, Ada Woo, Andrew Wiley and Thanos Patelis. When the pandemic forced the cancellation of the Annual Meeting, they pivoted quickly to transform the in-person program into a series of webinars that took place over the summer and fall of 2020. We learned a lot from their experience and adopted the new innovative session format – the Research Blitz – that they originated.

And as with the traditional in-person meeting, we have collaborated with many of you to prepare a program that we hope offers opportunities to learn, grow, connect, and celebrate some incredible work and achievements in the field. We are so excited about the conference and hope you enjoy it, even if we are socially distant (for now)!

Susan Davis-Becker and Leslie Keng
2021 NCME Annual Meeting Co-Chairs

Table of Contents

For technical assistance during the conference, please contact ncme@talley.com

Cover Image: Bunker Hill Bridge – Boston (unknown photographer)

NCME Leadership & Conference Team

Board of Directors

President	*Ye Tong, Pearson*
Vice President	*Derek Briggs, University of Colorado Boulder*
Past President	*Stephen Sireci, University of Massachusetts*
Members	*Ellen Forte, edCount* *Debbie Durrence, Gwinnett County Public Schools* *Andrew Ho, Harvard Graduate School of Education* *Sharyn Rosenberg, National Assessment Governing Board* *Michael Walker, The College Board* *Howard Everson, SRI International & City University of New York*

2021 Annual Meeting Chairs

Program Chairs	*Susan Davis-Becker, ACS Ventures* *Leslie Keng, Center for Assessment*
Training Chair	*Sarah Quesen, Pearson*

Fitness Run/Walk Directors

Jill R. van den Heuvel, Alpine Testing Solutions
Katherine Furgol Castellano, Educational Testing Service
Brian F. French, Washington State University

Publication Editors

Journal of Educational Measurement	*Sandip Sinharay, Educational Testing Service*
Educational Measurement: Issues and Practice	*Deborah J. Harris, University of Iowa*
ITEMS	*Andre Rupp, Mindful Measurement*
NCME Book Series	*Kadriye Ercikan, ETS*
NCME Newsletter	*Art Thacker, HumRRO*
NCME Website	*Matthew Gaertner, WestEd* *Brian Leventhal, James Madison University*

NCME 2021 Proposal Reviewers

Terry Ackerman
Anthony Albano
Benjamin Andrews
Alvaro J. Arce
Erin Banjanovic
Michelle Derbenwick Barrett
Michael Beck
Kirk Becker
Yufeng Berry
Pavneet Kaur Bharaj
Mustafa Kuzey Bilir
Michelle Boyer
Willard Christopher Brandt
Chad W. Buckendahl
Tiago A. Caliço
Luciana Cancado
Roti Chakraborty
Dandan Chen
Michelle Y. Chen
Yi-Hsin Chen
Hye-Jeong Choi
Seungwon Chung
Gregory Cizek
Kimberly Colvin
Jenna Copella
Nathan Dadey
Ted Daisher
Jennifer Davis
Laurie Davis
Susan Davis-Becker
Juan Manuel D'Brot
Onur Demirkaya
Nina Deng
Robert Dolan
Chris Domaleski
Bryan R. Drost
Denis Dumas
Ozge Ersan
Fen Fan
Rich Feinberg
Zachary Feldberg
Steve Ferrara
Leah Feuerstahler
Anthony D. Fina
Steve Fitzpatrick
Claudia Flowers
Brett P. Foley
Yanyan Fu
Robert Thomas Furter
Tracy Gardner
Yuan Ge
Abolfazl Ghasemi
Melissa L. Gholson
Brian Gong
Xaviera Gonzalez-Wegener
Guher Gorgun
Raman Grover
Gulsah Gurkan
Danielle Guzman-Orth
Heather Anne Handy
Qiwei He
Yong He
Briana Hennessy
Thomas P. Hogan
Timothy Scott Holcomb
Minju Hong
Jeffrey Hoover
Kristen Huff
Anne Corinne Huggins-Manley
Yvette Yvette Jackson
Unhee Ju
Hyun Joo Jung
Kwanghee Jung
Kate PS Kahoa
Yusuf Kara
Hacer Karamese
Gamze Kartal
Daniel Katz
Russell Keglovits
Leslie Keng
Justin L. Kern
Eunbee Kim
Jungnam Kim
Stella Kim
Young Yee Kim
Youngmin Kim
G. Gage Kingsbury
Emma M. Klugman
Jennifer Kobrin
Audra Kosh
Kevin Krost
Hollis Lai
Jon Lehrfeld
Brian C Leventhal
Xin Li
Zhen Li
Hwanggyu Lim
Chunyan Liu
Huan Liu
Na Liu
Susan Lottridge
Chang Lu
Ru Lu
Yong Luo
Ye Ma
Jaime Malatesta
Katerina Marcoulides Barbour
Scott Marion
Martha McCall
Janet Mee
Jing Miao
Lora Monfils
Scott Monroe
Kristin M. Morrison
Aaron Myers
Kyle Nickodem
Maura O'Riordan
Justin Paulsen
Michael R Peabody
Marianne Perie
Maria Potenza
Sonya Powers
Jia Quan
Sarah Quesen
Stanley N Rabinowitz
Michael Ralph
Vimal V Rao
Aileen Reid
Kelly Rewley
Fusun Sahin
Natalie Schelling
Madeline Schellman
Deborah Schnipke
Matthew Schultz
Benjamin R. Shear
Lijun Shen
Mark David Shermis
David Shin
Pooja Shivraj
Dubravka Svetina Valdivia
Nadine Talbot
Tony Thompson
W. Jake Thompson
Yeow Meng Thum
Lissette Tolentino
Ye Tong
Anna Topczewski
Sarah Linnea Toton
Emily Karen Toutkoushian
Anne Traynor
Jon S. Twing
Esther Ulitzsch
Chun Wang
Songtao Wang
Weimeng Wang
Yibo Wang
Zhaoyu Wang
Kelley Wheeler
Andrew Wiley
Anita Wilson
Ada Woo
Tong Wu
Yi-Fang Wu
Jiawei Xiong
Jiajun Xu
Ji Seung Yang
Seyma Nur Yildirim-Erbasli
Hanwook Yoo
Fabian Zehner
Mingqin Zhang
Mo Zhang
Mingying Zheng

Training Sessions

Full Day Training Sessions

Bayesian Networks in Educational Assessment

Part 1: Friday, June 4th | 9:00 AM - 1:00 PM EST
Part 2: Friday, June 4th | 1:30 PM - 5:30 PM EST
Duanli Yan, ETS; Russell G Almond, Florida State University; Diego Zapata-Rivera, ETS

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, while the second part will concentrate on model building and recent developments in the field.

Statistical Learning of Process Data: Methods, Software, and Applications

Part 1: Friday, June 4th | 9:00 AM - 1:00 PM EST
Part 2: Friday, June 4th | 1:30 PM- 5:30 PM EST
Jingchen Liu, Columbia University; Xueying Tang, University of Arizona; Susu Zhang, University of Illinois at Urbana

This full-day workshop introduces a selection of statistical learning methods for analyzing process data, that is, log data from computer-based assessments. Covered topics include (1) data-driven methods for extracting features from response processes; (2) sequence segmentation and subtask analysis with neural language modelling; (3) introduction to ProcData, an R package for process data analysis; and (4) applications of process features to practical testing and learning problems, including scoring, differential item functioning correction, computerized adaptive testing, and adaptive learning. Mode of instruction will be a blend of presentations, for topics (1) and (2), and concrete illustrations in R, for topics (3) and (4). Intended audience are researchers and practitioners interested in data-driven methods for analyzing process data from assessments and learning environments. To fully engage in the hands-on activities, familiarity with R and RStudio is expected. Running the ProcData package requires installation of R, Rcpp, and Python. Installation instructions and support will be provided. Participants are expected to bring their own laptop with Windows or Mac operating system. By the end of the workshop, participants are expected to get a composite picture of process data analysis and know how to conduct various analyses using the ProcData package.

Split Full Day Training Sessions

Creating Custom Interactive Applications with R and Shiny

Part 1: June 4th | 1:30 PM - 5:30 PM EST
Part 2: June 7th | 1:30 PM - 5:30 PM EST
Christopher Runyon, National Board of Medical Examiners; Joshua Goodman, National Commission on Certification of Physician Assistants; Marcus Walker, National Commission on Certification of Physician Assistants

This session explores the use of R and the Shiny package for creating unique statistical apps. In many testing, commercial, and academic contexts, there is a need for specialized statistical apps for custom tasks and analyses. Many of the commercially available programs are offered in a one-size-fits-all format, and thus often lack the flexibility needed across multiple contexts. Shiny is a free, open-source resource that can be used to build applications that can be developed and maintained by persons with only a modest level of R programming skill. These apps can be hosted on a webpage or deployed as standalone executable files, and end users of such apps do not need to know any R programming to successfully use them.

Using psychometric tasks as motivating examples, we guide session participants through building a simple app in Shiny. After teaching the foundations of a Shiny program, we expand to showcase some of the advanced capabilities of Shiny use, including generating reports and building standalone executable programs. Participants should have at least a moderate level of R programming ability. More advanced R programmers will still benefit from Shiny information that goes well beyond “hello world” examples often found on Shiny resource pages.

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

Part 1: June 7th | 9:00 AM – 1:00 PM EST
Part 2: June 8th | 9:00 AM – 1:00 PM EST
Jimmy de la Torre, University of Hong Kong; Wenchao Ma, University of Alabama

The primary aim of the workshop is to provide participants with the necessary practical experience to use cognitive diagnosis models (CDMs) in applied settings. Moreover, it aims to highlight the theoretical underpinnings needed to ground the proper use of CDMs in practice.

In this workshop, participants will be introduced to a proportional reasoning (PR) assessment that was developed from scratch using a CDM paradigm. Participants will get a number of opportunities to work with PR assessment-based data. Moreover, they will learn how to use GDINA, an R package developed by the instructors for a series of CDM analyses (e.g., model calibration, evaluation of model appropriateness at item and test levels, Q-matrix validation, differential item functioning evaluation). To ensure that participants understand the proper use of CDMs, the theoretical bases for these analyses will be discussed.

The intended audience of the workshop includes anyone interested in CDMs who has some familiarity with item response theory (IRT) and R programming language. No previous knowledge of CDM is required. By the end of the session, participants are expected to have a basic understanding of the theoretical underpinnings of CDM, as well as the capability to conduct various CDM analyses using the GDINA package.

Full Day Training Sessions (4-hour asynchronous, 4-hour session)

Using Stan for Bayesian Psychometric Modeling

June 7th | 1:30 PM - 5:30 PM EST

Yong Luo, ETS; Manqian Liao, Duolingo

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

Morning Half-Day Sessions

A Visual Introduction to Computerized Adaptive Testing

June 8th | 9:00 AM - 1:00 PM
Yuehmei Chien, College Board; David Shin, Pearson

The training will provide the essential background information on operational computerized adaptive testing (CAT) with an emphasis on CAT components (including ability estimation, item exposure control and content balancing methods--weighted penalty model and shadow tests) and CAT simulation. Besides the traditional presentation through slides, this training consists of hands-on demonstrations of several key concepts, with visual and interactive tools and a CAT simulator.

Practitioners, researchers, and students are invited to participate. A background in IRT is recommended. Participants should bring their own laptops and item pools in CSV file format, as they will access the tools that were designed to help the participants understand important CAT concepts and visualize the results. Installation instruction of the tools will be provided via email prior to the conference. Upon completion of the workshop, participants are expected to have 1) a broader picture about CAT; 2) a deeper understanding of the fundamental CAT techniques; 3) appreciation of the visual techniques used to analyze and present results in an intuitive and pleasing way.

Addressing the Data Challenges from Next-generation Assessments: Data Science Upskilling for Psychometrician

June 8th | 9:00 AM - 1:00 PM
Oren Livne, ETS; Jiangang Hao, ETS

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

An Overview of Operational Psychometric Work in Real World

June 8th | 9:00 AM - 1:00 PM
Hyeon-Joo Oh, ETS; JongPil Kim, Riverside Insights; Jinghua Liu, Enrollment Management Associates; Sarah Quesen, Pearson; Hanwook Yoo, ETS

An overview of the psychometric work routinely done at various testing organizations will be presented in this training session. The training session will focus on the following topics: (1) outline of operational psychometric activities across different testing companies, (2) hands-on activities to review item and test analyses output, (3) hands-on activities to review equating output, and (4) overview of computerized adaptive testing (CAT) and multi-stage testing (MST) and hands-on activities. If time allows, there will be a brief discussion session regarding factors that affect operational psychometric activities in the CAT ad MST environment. 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, and etc.) will present various topics related to processes in an operational cycle.

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

June 8th | 9:00 AM - 1:00 PM

Seung W. Choi, University of Texas Austin; Sangdon Lim, University of Texas Austin

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

Python, Machine Learning, and Applications - A Gentle Introduction

June 8th | 9:00 AM - 1:00 PM

Zhongmin Cui, CFA Institute

Machine learning is getting popular in recent years. Its applications span a vast range: from agriculture to astronomy, from business to biology, from communication to chemistry, from data mining to dentistry, from education to economy; the list goes on. The interest in machine learning continues growing as indicated by related presentations and publications. The goal of this lecture-style training is to provide a gentle introduction on this topic. Although other languages are available for machine learning, Python will be introduced as a starter in this training. The main course has two dishes, supervised machine learning and unsupervised machine learning. Dessert samples of using machine learning in educational measurement research conclude the training. Participants do not need to have any experience in machine learning or Python. Upon completion, participants are expected to have a general idea of machine learning and know how to use Python on a simple machine learning project. Participants do not need to bring their laptops or install software; the training will be as gentle as possible so that it is tasty to a broad audience. Having said this, following an example with a laptop near the end of the training would make the dessert taste sweeter.

ReCo: A Shiny App for Automatically Coding Short Text Responses in Assessments

June 8th | 9:00 AM - 1:00 PM

Fabian Zehner, DIPF | Leibniz Institute for Research and Information in Education, Centre for International Student Assessment, ZIB); Nico Andersen, DIPF | Leibniz Institute for Research and Information in Education

In this training session, participants will learn to use the ReCo shiny app (Automatic Text Response Coder) for automatically coding text responses in assessments. For example, this can be used for scoring constructed responses as correct or incorrect. The session will start with an introduction to the employed methodology (i.a., Latent Semantic Analysis, classification and its evaluation) but will have its focus on hands-on activities. Participants will use a graphical interface in R for automatically coding text responses from an English response data set. Participating assessment developers, practitioners, as well as researchers will be empowered to automatically code constructed responses in their own assessments.

Afternoon Half Day Sessions

Bridging Research and Practice by Examining the Consequences of Assessment Design and Use

June 8th | 1:30 PM - 5:30 PM

David Slomp, University of Lethbridge; Maria Elena Oliveri, Buros Testing Center, University of Nebraska-Lincoln

Participants will learn how to systematically examine the consequences of assessment design and use for both classroom and large-scale assessment programs.

Participants will be introduced to two approaches—Integrated Design and Appraisal Framework (Slomp, 2016) and Theory of Action (Bennett, 2010)—for integrating attention to consequences into the design and appraisal of assessment programs. The IDAF approach provides a taxonomy for considering questions of fairness, validity and reliability in an integrated fashion that highlights the intended and unintended consequences of decisions made at each stage of an assessment's design and use. The Theory of Action (ToA) model applies logic models to the articulation and testing of claims regarding both how program information is used, and the impact using this information has on individuals or organizations.

An overview of the literature on the consequences of assessment design and use will be provided. Participants will then be guided through two case studies illustrating the application of the IDAF and ToA models. Participants will then work collaboratively on building a plan of action, extrapolated from these frameworks, that they will apply to a third case study.

Computerized Multistage Testing: Theory and Applications

June 8th | 1:30 PM - 5:30 PM

Duanli Yan, ETS; Alina A. von Davier, DuoLingo; Kyung (Chris) Han, Graduate Management Admission Council

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

Modeling Writing Process Using Keystroke Logs

June 8th | 1:30 PM - 5:30 PM

Mo Zhang, ETS; Hongwen Guo, ETS; Xiang Liu, ETS

In this half-day workshop, participants will have an opportunity to learn about and analyze a newer type of the educational data that is being progressively used in writing research; namely, the keystroke logs collected during the writing process. Information contained in the keystroke logs goes much beyond a holistic evaluation on the written product. From the keystroke logs, one may identify, for example, whether a writer had trouble with retrieving words, edited what was written before the submission, or spent sufficient time and effort on the task. As much as the opportunities and potential applications given by this type of timing and process data, it also poses many challenges to researchers and practitioners, which includes construct-relevant evidence identification from the logs, evidence extraction/feature engineering, and statistical treatment and modeling of such complex data. Students and professionals in the areas of writing research and educational measurement are invited. The format of this workshop will be a mix of lecture-style presentation, hands-on data analyses, and group discussion. Some background on statistical analyses will be preferred. Sample R codes for applying Markov or semi-Markov process and other graphical models will be provided. Participants should bring personal laptops with the statistical software R installed.

Principles and Methods in Psychometric Evaluation of Educational Assessments

June 8th | 1:30 PM - 5:30 PM

Louis Roussos, Cognia; Han Yi Kim, ACT; Liuhan, Sophie Cai, Cognia

Dr. Louis Roussos has 15 years of experience in evaluating test forms from a variety of assessment programs. He is in the early stages of writing a textbook and will share the "tricks of the trade" he has learned, including guiding principles, methods that flow from these principles, and a variety of real-life examples. Special consideration will be given to importance of communication and complex decision making. The session will entail a mixture of lecture, dialogic learning through interactive discussion and sharing of experiences by participants, methods demonstration, and practical exercises in which participants implement the principles and methods. The exercises will result in constructive feedback to both the presenters and participants.

Using SAS for Monte Carlo Simulation Studies in Item Response Theory

June 8th | 1:30 PM - 5:30 PM

Brian Leventhal, James Madison University; Allison Ames Boykin, University of Arkansas

Data simulation and Monte Carlo simulation studies are important skills for researchers and practitioners of educational measurement, but there are few resources on the topic. This four-hour workshop presents the basic components of Monte Carlo simulation studies (MCSS). Multiple examples will be illustrated using SAS including simulating total score distribution and item responses using the two-parameter logistic IRT, bi-factor IRT, and graded response model. Material will be applied in nature with considerable discussion of SAS simulation principles and output.

The intended audience includes researchers interested in MCSS applications to measurement models as well as graduate students studying measurement. Comfort with SAS base programing and procedures will be helpful. Participants are not required to have access to SAS during the session. The presentation format will include a mix of illustrations, discussion, and hands-on examples.

As a result of participating in the workshop, attendees will: 1) Articulate the major considerations of a Monte Carlo simulation study, 2) Identify important SAS procedures and techniques for data simulation, 3) Adapt basic simulation techniques to IRT-specific examples, and 4) Extend examples to more complex models and scenarios.

Using School-Level Data from the Stanford Education Data Archive

June 7th | 1:30 PM - 5:30 PM

Sean Reardon, Stanford University; Andrew Ho, Harvard Graduate School of Education; Benjamin R. Shear, University of Colorado Boulder; Erin Michelle Fahle, St. John's University

The Stanford Education Data Archive (SEDA) is a growing, publicly-available database of academic achievement and educational contexts. The nationally-comparable achievement data is based on roughly 500 million standardized test scores for students in nearly every U.S. public school in third through eighth grade from the 2008-09 through 2017-18 school years. Initially, SEDA included only estimates of school district and county-level achievement. Subsequently, the data were expanded to include estimates of average school-level achievement and were made accessible to a broader audience through a new, interactive website, The Educational Opportunity Project Data Explorer.

This workshop is intended to introduce researchers of all levels, practitioners, and policymakers to the school-level SEDA achievement and context data. We will provide an overview of both the contents of the SEDA database and the statistical and psychometric methods used to construct the database. The workshop will include presentations by the instructors and hands-on activities designed to help users engage directly with the school-level data. All attendees should have a computer accessible during the training in order to engage in the activities. Attendees who are interested in using the data for research purposes should have statistical software (e.g., R or Stata) installed on their computers.

A Case Study in Measurement Practice and The Public Perception

11:00 to 12:30 pm – Organized Discussion

As measurement professionals, we prioritize aspects of test validity, fairness, and appropriate score use in the development and administration of our exams as well as the interpretation of test scores. Many highly trained psychometricians and researchers spend many hours on the critical work of establishing, examining, and improving these aspects of our assessments. Both the related procedural work as well as the outcomes, analyses, and findings are not easily communicated to the public at large and therefore present opportunities for clearer communication and more compelling ways of sharing what we do and what we know about our assessments. In this session, we will share information around three key assessment issues: (1) Fairness; (2) Validity and value; and (3) Use for accountability. We will review how each of these areas is addressed, operationalized, and informed by research, practice, and broader implications within large testing organizations, using many examples from the SAT. After sharing information on each of those areas, we will hear commentary from experts in those three areas to identify and highlight avenues for improved public understanding of research and practice and consider additional work we could or should be doing in those areas.

Session Organizer:

Emily Shaw, College Board

Presenters:

Rebecca Zwick, Educational Testing Service
Michael E. Walker, Educational Testing Service
Brent Bridgeman, ETS
Jonathan Beard, College Board
Ellen Forte, edCount, LLC

(Invited Session) Lessons about the modeling and measurement of human abilities

11:00 to 12:30 pm – Organized Discussion

Both educational measurement and quantitative psychology share an origin story in the pioneering work of Charles Spearman near the turn of the 20th century. Spearman's methodological contributions, which evolved into what we today might recognize most readily as classical test theory and latent variable modeling, were in service of his two-factor theory of intelligence. Many of the lessons from the longstanding debate over Spearman's g, and his method for "measuring" it, can be reflected in innovative ideas emerging out of quantitative psychology over the past decade. In this session, Derek Briggs will provide some historical and conceptual context for Spearman's theory and the controversy this produced about the structure of human abilities. Wes Bonifay, Li Cai and Riet van Bork will then present on some new ideas for how human abilities can be modeled, measured and interpreted. The session will conclude with a discussion about insights history and modern innovations suggest for improvements in the design, analysis and use of educational tests.

Session Organizer:

Derek Briggs, University of Colorado

Presenters:

Derek Briggs, University of Colorado
Wes Bonifay, University of Missouri
Li Cai, UCLA
Riet van Bork, Center for Philosophy of Science, University of Pittsburgh

Embedded Standard Setting: Research & Advances

2:00 to 3:30 pm - Coordinated Paper Session

The traditional item-based standard-setting workshop is a common component of the assessment lifecycle typically conducted following the first operational test administration. In contrast, Embedded Standard Setting (ESS; 2020, Lewis & Cook), a relatively new methodology situated in a Principled Assessment Design framework, is embedded in processes that occur throughout the assessment lifecycle. When the requirements are met, ESS cut scores emerge organically and algorithmically by optimizing the consistency of hypothesized item alignments and item difficulty. Lewis & Cook (2020) assert that if the requirements of ESS are met then the traditional standard-setting workshop is redundant at best and contradictory at worst. This session extends the basic exposition of the ESS methodology published by Lewis & Cook (2020) with research and advances that:

- explore how the ESS response probability can be optimized,
- compare an alternate ESS-cut-score algorithm to that proposed by Lewis & Cook (2020),
- investigate the magnitude of correlation between item alignment and item difficulty with respect to the efficacy of ESS,
- leverage an ESS enhancement to a popular item-based standard-setting method—ID Matching—to reduce panelists' cognitive load, and
- explore the application of ESS methods to add value to a mature testing program.

Session Organizer & Chair:

Daniel Lewis, Creative Measurement Solutions LLC

Participants:

Optimal Response Probabilities in Embedded Standard Setting

Robert Cook, Cognia; Daniel Lewis, Creative Measurement Solutions LLC

A Comparison of Two ESS Cut Score Estimation Algorithms

Daniel Lewis, Creative Measurement Solutions LLC; Sooyong Lee, University of Texas at Austin

The Alignment-Data Coherence Criterion for Embedded Standard Setting

Jing Chen, NWEA; Daniel Lewis, Creative Measurement Solutions LLC; Robert Cook, Cognia

Embedded ID Matching: Applying ESS to Reduce ID Matching Panelist Cognitive Load

Christina Schneider, NWEA; Daniel Lewis, Creative Measurement Solutions LLC

Practical Applications of Embedded Standard Setting Methods in a Mature Testing Program

Rachel R Kachchaf; Leslie Pearlman, ETL LLC; Daniel Lewis, Creative Measurement Solutions LLC

Discussant:

Steve Ferrara, Cognia

Psychometric Challenges and Potential Solutions for Educator Testing in Pandemic Environment

2:00 to 3:30 pm - Coordinated Paper Session

COVID-19 disrupted day-to-day operational scoring and reporting practices of licensure and certification testing programs. The coordinated session discusses issues and potential solutions to the change in candidate demographics and educational characteristics resulting from implementing economic relief measures in Florida in response to COVID-19. The session brings five presentations covering issues and short-term solutions on development and reporting practices for the Florida Teacher Certification Examinations (FTCE), such as field testing, equating, and scoring. The presentations use archival data set collected before and during the pandemic to investigate the repercussions and efficacy of short-and-medium term solutions on FTCE score uses and interpretations.

Session Organizer:

Alvaro J. Arce, Pearson

Participants:

Free Testing and Pandemic Impact on Field Testing Outcomes

Lauren White, Florida Department of Education

Effects of Change in Intended Test-Taker Population on Pre and Post Equating Outcomes

Alvaro J. Arce, Pearson

Holistic Scoring Challenges Caused by Stay-at-Home Order and Potential Solutions

Sarah Underwood, Florida Department of Education

Free Testing Impact on Educator Assessments and Examinee Performance in Pandemic Environment

Leah Kaira, Pearson

Pandemic and Free Testing Impact on Psychometric Properties of Educator Testing

Suleyman Olgar, Florida Department of Education

Unpacking Cognitive Complexity: What is it and Why is it so Hard?

11:00 to 12:30 pm - Organized Discussion

Every item that is used in large-scale k-12 assessment must be coded for depth of knowledge. The DOK is used in form construction to ensure that a breadth of cognitive demands is represented on the assessment. Alignment studies review DOK for items and learning standards to evaluate match. However, there is a call in the field for a different lens through which to view cognitive complexity and to more explicitly incorporate cognitive complexity into assessment design and score interpretation. This call reflects a number of persistent challenges with respect to cognitive complexity: conceptualization, measurement, and implementation. Join scholars and practitioners for a lively discussion of these challenges and potential solutions.

Session Organizer:

Kristen Huff, *Curriculum Associates*

Presenters:

Steve Ferrara*, Cognia*
Ellen Forte*, edCount, LLC*
James Pellegrino*, University of Illinois at Chicago*
Christina Schneider*, NWEA*

(Invited Session) Using Longitudinal Assessment to Support Professional Development
11:00 to 12:30 pm - Organized Discussion

In recent years, several medical certification boards have begun to design and incorporate longitudinal assessments (i.e., relatively small batches of items administered at regular intervals over an extended period of time) into their continuing certification programs. One of the primary drivers behind this movement has been the desire to improve the formative feedback provided to participants for the purposes of guiding and supporting ongoing education and professional development. In this session, representatives from four certifying organizations will share their experiences with longitudinal assessments. Collectively, the presenters will describe design elements intended to guide or promote continuing education efforts. The presenters will also share data that illustrates how participants are engaging with and performing on these types of assessments, which will provide insight into whether longitudinal assessments are meeting their intended objectives from a formative perspective.

Session Organizer & Chair:
Andrew Dwyer, *American Board of Pediatrics*

Presenters:
Robert C. Shaw, National Board for Respiratory Care
Huaping Sun, American Board of Anesthesiology
Ying Du, American Board of Pediatrics
Robert Brucia, American Board of Pediatrics
Ting Wang, American Board of Family Medicine

Procedures for establishing and evaluating linkages between scores collected in different modes
2:00 to 3:30 pm – Coordinated Paper Session

The collection of papers in this session will discuss linking methodologies commonly used to enable proper comparison of results from digital tests and the corresponding paper tests. Selecting the appropriate linking method starts with understanding the testing program's needs, its practical feasibility, as well as constraints in carrying out different linking designs. Small-scale research studies and/or field trials are often used to empirically validate the method of choice. After the operational data are collected, psychometric assumptions associated with the selected method should be checked and the score comparability issue should be evaluated comprehensively. This coordinated session will provide a general introduction on different linking methods between testing modes and discuss how to choose among these methods in practice, by referring to empirical linking experience. State assessments and large scale educational survey assessments will be used as examples. The objective is to share the technical knowledge developed across these testing programs as well as their substantive findings to assist practitioners in better designing their linking studies for appropriate mode comparison. The session will include discussion from a leading expert in the field from technical and practical perspectives.

Session Organizer:
Nuo Xi, VIPKID

Participants:
Design considerations for linking large scale survey assessments across modes
Yue Jia, Educational Testing Service; Nuo Xi, VIPKID
Common population linking method used in NAEP digital transitions
Nuo Xi, VIPKID; Paul Adrian Jewsbury, Educational Testing Service
IRT model extensions for modeling mode effects in PISA 2015
Matthias Von Davier, NBME; Lale Khorramdel, Boston College

Discussant:
Tim Moses, College Board

Ethics & STEM Assessments: Content modeling, construct mapping, psychometric models, mitigating bias
2:00 to 3:30 pm - Coordinated Paper Session

Project Lead The Way (PLTW), a national project-based STEM curriculum provider, partnered with the University of Kansas (KU) to replace prior assessments, which were multiple choice, unidimensional, and measured only subject-matter knowledge, with a new type of assessment that measured ethical reasoning as well as subject-matter knowledge in the context of the STEM content. The aim was to improve the relevance and validity of the assessments based on the research-based prediction that this approach would mitigate historical gender bias in STEM assessment while simultaneously establishing a new paradigm in assessment design. The design effort included careful evaluation of the curriculum frameworks, drafting of achievement level descriptors, creation of aligned test blueprints, and the design and development of tasks and task models using an evidence-centered design approach. In the 18-19 school year, approximately 340,000 End-of-Course (EoC) assessments across 14 courses were administered to students in all 50 states. This Coordinated Paper Session discusses the approaches utilized to develop valid assessments that measured ethical reasoning as well as subject-matter knowledge and highlights how careful selection of what is included in an assessment and the evidence underlying such selection can have profound effects on fairness and bias across gender, racial, and ethnic lines.

Session Organizer:
Michelle Gough, EdMetric, LLC

Participants:

Construct mapping to elicit expert opinions, measuring ethical reasoning in end-of-course assessments
Richard Patz, University of California, Berkeley; Neal Kingston, University of Kansas; Michelle Gough, EdMetric, LLC

Daubert-Style Evidence Centered Design: Blueprint Design, Content Modeling, and Standard Setting
Michelle Gough, EdMetric, LLC; Emily Richardson, Project Lead The Way; Karla Egan, EdMetric, LLC

Wrestling with multidimensionality in the measurement of ethical reasoning in STEM assessment
Neal Kingston, University of Kansas; Richard Patz, University of California, Berkeley; Ashley Williams, Bioplicity; Karla Egan, EdMetric, LLC

Investigating demographic-specific achievement and correlations between ethical reasoning and other transferable skills
Ashley Williams, Bioplicity; Richard Patz, University of California, Berkeley; Michelle Gough, EdMetric, LLC; Neal Kingston, University of Kansas

Discussant:
Michael Joseph Smith, University of Virginia

(SIGIMIE Session) Current Challenges in Large-scale Assessment and Responses/Innovations

11:00 to 12:30 pm - Organized Discussion

In this panel, we aim to discuss the current challenges in large-scale assessments (LSAs) and how different agencies may respond to them by adopting various approaches and innovations; and in turn, how those approaches vary across state, national, and international assessments; high- and low-stakes; and across operational and academic works. For instance, in international LSAs, priorities for the OECD may involve addressing challenges related to the relevance for policies of dozens of different participating educational systems or measures and expansions to systems that are economically different from one another; these considerations are perhaps less important for a state department of education. Similarly, while issues of fairness in testing are considered important across all of educational assessment, how they manifest themselves may 'look' and be addressed differently. There might also be differences in what we believe the purpose of the assessment is, and what should our goals be with respect to that assessment. Our panel experts bring experiences, knowledge, and understanding of the role of assessments and how the current challenges might be responded to (via innovation) from a variety of different levels.

Session Organizers:

Dubravka Svetina Valdivia, Indiana University
Leslie Rutkowski, Indiana University

Presenters:

Andrew Ho, Harvard Graduate School of Education
Kristen Huff, Curriculum Associates
Charity Flores, Indiana Department of Education
Matthias Von Davier, NBME

Fair and Valid Assessment of ELs with the Most Significant Cognitive Disabilities

11:00 to 12:30 pm - Coordinated Paper Session

This coordinated session addresses the fair and appropriate assessment of ELs with the most significant cognitive disabilities. The assessment of ELs with the most significant cognitive disabilities is a federal requirement. A challenge in meeting federal requirements and ensuring valid, fair, and equitable assessment and accountability includes understanding the characteristics of the student population and the contexts in which they learn, defining the language domains appropriately, and articulating reasonable expectations as well as appropriate conditions vis-a-vis demonstrations of what they know and can do. Published research on this student population is limited. However, there are a few recent efforts that systematically examine these students, their linguistic and cultural capabilities and backgrounds, and their instructional and assessment experiences, to inform fair and valid assessment practices. The presenters, who have been involved in these recent efforts, will report on research and development relevant to the English language proficiency assessment of ELs with the most significant cognitive disabilities. Policy, heuristics for ensuring fair and valid measures, and research findings that can inform principled approaches to assessment design and development for these students will be presented. Information shared in this session is intended to be useful to assessment designers and developers, researchers, and policymakers.

Session Organizer:

Edynn Sato, Sato Education Consulting LLC

Participants:

Alternate English Language Proficiency Assessment: Federal Requirements, Heuristics, and Promising Practices
Deborah Spitz, U.S. Department of Education

The Individual Characteristics Questionnaire: Understanding ELs with the Most Significant Cognitive Disabilities
Laurene Christensen, WIDA at the Wisconsin Center on Education Research

Operationalizing Language Domains for ELs with Significant Cognitive Disabilities: Designing Fair and Valid Measures
Edynn Sato, Sato Education Consulting LLC

Developing Item Templates for Alternate Assessments of English Language Proficiency
Phoebe C Winter

Discussant:

Christopher Rivera, East Carolina University

(SIGIMIE Session) Scaling, Linking, & Equating Du Jour: A Discussion with Experts

2:00 to 3:30 pm - Organized Discussion

This past year, NCME initiated Special Interest Groups in Measurement in Education (SIGIMIEs) to increase member retention and engagement. The Contemporary Issues in Scaling, Linking, and Equating (SLE) SIGIMIE currently has 108 members. As chairs of the SLE SIGIMIE, we are proposing an organized discussion for the NCME 2021 Annual Meeting that will feature a panel of SLE experts discussing various topics raised by SLE SIGIMIE members.

Session Organizers:

Stella Kim, University of North Carolina at Charlotte
Jaime Malatesta, Graduate Management Admission Council

Presenters:

Michael Kolen, University of Iowa
Marie Wiberg, Department of Statistics, USBE
Tim Moses, College Board
Jorge Gonzalez, Pontificia Universidad Catolic

Discussant:

Robert Brennan, University of Iowa

(Invited Session) Advancing Women in Measurement: Barriers and Opportunities

2:00 to 3:30 pm - Organized Discussion

This session will provide a venue to discuss the status of women in educational measurement, barriers to equal representation in leadership positions, and opportunities for overcoming those barriers. Women in measurement have a long history of leadership and impact. However, there continue to be significant gender and racial disparities in who holds the most influential and highly-regarded positions in our field. The session will open with a presentation from Susan Lyons highlighting the underrepresentation of women—and female minorities in particular—in positions of leadership in our field. Susan will offer paths forward for understanding and dismantling systems of oppression that perpetuate those inequities. To foster discussion on these issues, Jenn Dunn will moderate a panel of women who have excelled at the highest levels: Ye Tong, Jennifer Randall, and Ellen Forte. Our panelists represent diverse cultural backgrounds and have each forged distinguished career paths in different sectors of our field. We will then hear commentary from our discussant, Kadriye Ercikan, who will provide insight and perspective on the issues and opportunities discussed. The last fifteen minutes will be reserved for interaction with participants who will be invited to share their own experiences, reflections, and questions for the speakers.

Session Organizer:

Susan Lyons, Women in Measurement, Inc.

Chair:

Jennifer Dunn, Questar

Presenters:

Susan Lyons, Women in Measurement, Inc.
Ye Tong, Pearson
Jennifer Randall, University of Massachusetts
Ellen Forte, edCount, LLC
Kadriye Ercikan, Educational Testing Service

Psychometrics for Digital-First Assessments: The Duolingo English Test Application

11:00 to 12:30 pm - Coordinated Paper Session

With the growth of digital technology and advances in automated test development tools, ranging from automated item generation to automated scoring, opportunity has come to develop innovative forms of technology-based assessments. This symposium offers an overview of how these advanced technologies support a valid and reliable digital-first test, the Duolingo English Test (DET), which is designed to be accessible anywhere and anytime. The four selected papers cover roughly the main parts of the test development process: The first paper provides an overview of the DET item development and evaluation procedure. The second paper introduces a model that calibrates the item difficulty with both the operational response data and CEFR-labelled passages. The third paper presents a measurement model designed for the DET which has continuous response data. The fourth paper evaluates the fairness of the DET by assessing the differential item functioning across several subpopulations of interest. These studies also contribute to psychometric research from a perspective that is well beyond DET: They have been applied to a challenging scenario where items have little overlap across testing administrations and test takers by design. To conclude the session, a well-known expert in psychometrics will provide a discussion of these papers.

Session Organizer:
Manqian Liao, Duolingo

Chair:
Alina A von Davier, Duolingo

Participants:

Design, Development, and Evaluation of Duolingo English Test Items
Geoff LaFlair, Duolingo; Lauren Bilsky, Duolingo; Jesse Egbert, Northern Arizona University; Brent Burch, Northern Arizona University; Margaret Wood, Northern Arizona University

Improving Language Item Difficulty Estimation with BERT and Multi-Task Learning
Arya McCarthy, Johns Hopkins University; Kevin Yancey, Duolingo; Geoff LaFlair, Duolingo; Jesse Egbert, Northern Arizona University

The Duolingo English Test: Psychometric Considerations
Gunter Maris, ACT

Investigating Differential Item Functioning in Duolingo English Test
Manqian Liao, Duolingo; Geoff LaFlair, Duolingo

Discussant:
Stephen Sireci, University of Massachusetts Amherst

Assessing COVID-19 Impacts on Assessment and Learning using Star Interim Assessments

11:00 to 12:30 pm - Coordinated Paper Session

The coronavirus pandemic had led to unprecedented changes in the way we live our lives in the United States, including the ways in which students are educated. Beginning in Spring 2020, the education of students was dramatically interrupted due to the pandemic and continues into the 2020-2021 academic year. Given the unprecedented nature of the interruption to student education, there are numerous important and, at present, unanswered questions about what the impact is to student learning. Accurately determining the impact of the coronavirus pandemic on student learning is critically important as it has the potential to inform pedagogical and policy related decisions that can ameliorate the impact of the pandemic on student learning. At present, the highest quality, standardized, assessment data available to investigate impact on student learning comes from large scale, standardized, interim assessments. In this session we present four research studies based upon results from Renaissance Star interim assessments, the most widely administered interim assessment in the US. Each of the studies presents a distinct view on the impact of COVID-19 from its impact on the administration of assessments themselves to the impact on student learning as indicated by the assessment results.

Session Organizer:

Damian Betebenner, Center for Assessment

Participants:

How Did the Pandemic Impact Interim Assessment? An Analysis of 2020-21 Star Testing Patterns and Remote Testing Data

Katie McClarty, Renaissance; Adam Wyse, Renaissance

Estimating the Impact of Spring 2020 School Building Closures on Fall 2020 Performance and Student Readiness to Learn

Eric Stickney, Renaissance; Lindsay Haas, Renaissance

Evaluating Differential Impacts of School Building Closures on Fall 2020 Performance by Subgroup

Amanda Beckler, Renaissance; David Butz, Renaissance

Using SGPs from Star Assessments to Understand COVID-19 Learning Loss

Damian Betebenner, Center for Assessment

Discussant:

Derek Briggs, University of Colorado

Large-scale Educational Data Sets and the Ethics of their Monetization

2:00 to 3:30 pm - Organized Discussion

The increased use of commercial digital tools in public education has led to the creation of large data sets that contain detailed information about students' educational career. Such data sets, spanning applications from learning management systems to standardized assessments, hold potentially useful and actionable information that can support the development of targeted instruction, as well as more efficient and efficacious products. Discussions about these data have centered on questions of student privacy. However, data sets with the potential to track the educational progress of entire populations are of great economical value for the vendors of such products and services. In this roundtable we will discuss the ethical and practical implications of generating value, both monetary and scientific, from data that are the result of a public good. Stakeholders representing local and federal educational agencies, as well as educational software vendors, will debate questions of data ownership, access, and sharing as well as the ethical implications of generating value from data created in the pursuit of a public good: state sponsored education.

Session Organizer:

Tiago A. Caliço, American Institutes for Research

Presenters:

William Robert Buchanan, SAG Corp
Alina A von Davier, Duolingo
Emmanuel Sikali
Amelia Vance, Future of Privacy Forum

(Invited Session) Education literacy for psychometricians

2:00 to 3:30 pm - Organized Discussion

The NCME conference has traditionally been focused on the technical and practical challenges facing the educational measurement community. Over the past few years, multiple sessions have focused on assessment literacy, with discussions centered on training the broader educational community about critical measurement concepts. One aspect missing from these sessions is what NCME members can do to better understand the broader education community and what we can do to help make assessments better fit the practical realities educators must address every day. This session will aim to review and discuss what steps the NCME community can do to better understand how the assessments that we develop and deliver can be utilized in schools. The panel represents a broad range of experience in the educational measurement community and will discuss some key topics, such as: 1. How can the educational measurement community engage with the larger educational community so that both parties can learn and grow from one another? 2. What areas or topics in education would it be beneficial for the NCME community to know more about or better understand? 3. What can the NCME community do to allow the assessments we develop better fit within the broader educational context?

Session Organizer & Chair:

Andrew Wiley, ACS Ventures LLC

Presenters:

Stephen Sireci, University of Massachusetts Amherst
Susan Brookhart, Duquesne University
Rhonda True, Nebraska Department of Education
Debbie Durrence, Gwinnett County Public Schools

On the Assessment of Non-Cognitive Competencies in Licensure: Why, Whether, and How?

11:00 to 12:30 pm - Coordinated Paper Session

Speakers will share their research and opinions regarding the possibilities for and appropriate measurement of Non-Cognitive Competencies in licensure decisions. Non-Cognitive Competencies (NCCs) have long been included in employment testing, and recent research has demonstrated their success in admissions contexts. Perhaps as a result of these successes, there has been increasing interest in including NCCs in licensure decisions. However, there are important concerns around the inclusion and use of NCCs in licensure testing and downstream decisions, including questions of gameability, concerns regarding fairness toward subgroups of candidates, and questions around predictive validity, among others. The purpose of this session is to clarify the arguments in support of and against the inclusion of NCCs specifically within a licensure context, and to provide a forum for members of the profession to weigh in on this important and timely topic. Each speaker will make a presentation of his or her work and thinking related to the measurement of NCCs. The discussant will synthesize ideas across the presentations and invite conversation among the speakers, with audience participation welcomed.

Session Organizer:

Joanne Kane, National Conference of Bar Examiners

Participants:

Assessing Non-Cognitive Competencies in Legal Licensure: Lessons from Neighboring Fields

Joanne Kane, National Conference of Bar Examiners

The Use of Non-Cognitive Measures as a Component of Licensure and Certification Measures

Kurt Geisinger, University of Nebraska-Lincoln

Comingling Cognitive and Non-cognitive Competencies for Credentialing Exams: A Slippery Slope

Chad W. Buckendahl, ACS Ventures, LLC

Discussant:

Patrick Charles Kyllonen, ETS

(SIGIMIE Session) Building a Multidimensional Future: A Conversation on Big Data and Educational Measurement

11:00 to 12:30 pm - Organized Discussion

The most important factor for the increase in Big Data in education is the introduction of online/digital learning and assessment environments. These environments create enormous amounts of data on learner/examinee behaviors. The measurement field is essential for turning this data into smart data to benefit all aspects of assessment and learning. One of the main goals of Big Data in education is to leverage the availability of this data to inform the study of education. To help support this goal, the NCME Big Data in Educational Measurement SIGIMIE proposes a discussion panel session that highlights critical questions around Big Data and Educational Measurement and opens up the conversation to address the key challenges and opportunities in bridging Big Data research and practical use of Big Data. The proposed panel will discuss topics on: a) the role of Big Data in the field of education and educational measurement, b) Big Data collection, management, privacy, security, and ethics, and c) Big Data modeling/mining, among others. This panel will invite experts in the field to share their knowledge and experience on these various topics to increase the quality of conversations surrounding Big Data in the context of education and educational measurement.

Session Organizer:

Ruhan Circi, American Institutes for Research

Presenters:

Juanita Hicks, American Institutes for Research
Soo Lee, American Institutes for Research
Ryan Baker, University of Pennsylvania

(SIGIMIE Session) Debating the training of future measurement professionals

2:00 to 3:30 pm - Organized Discussion

Select measurement professionals representing universities, testing companies, K-12 education departments, and program evaluation consortia were surveyed on the skills they believe to be vital for success in their respective professions. In this session, we: (a) present the results from this survey in relation to graduate training; and (b) provide an opportunity for each participant to defend their own choices, challenge other groups' responses, and, perhaps, even change their own mind. In this debate style session, we encourage audience members to react to, and participate in, the spirited debate about exactly how to train measurement graduate students.

Session Organizers:

Brian C Leventhal, James Madison University
Joseph A. Rios, University of Minnesota
Corinne Huggins-Manley, University of Florida
Allison Ames Boykin, University of Arkansas

Presenters:

Deborah Bandalos, James Madison University
Peggy Jones, Pasco County (FL) Dist School
Kavita Mittapalli, MN Associates, Inc.
Jonathan Rubright, National Board of Medical Examiners
Vince Verges, Florida Department of Education

Guidelines for Technology-Based Assessments: An ITC and ATP Collaboration

2:00 to 3:30 pm - Coordinated Paper Session

In 2018, the International Test Commission and Association of Test Publishers agreed to partner on development of Guidelines for Technology-Based Assessments. This project has involved over 40 authors and reviewers across the Globe. In this session we discuss the Guidelines and how NCME members can contribute to the development process.

Session Organizer:

Stephen Sireci, University of Massachusetts Amherst

Participants:

Guidelines for Test Design

Craig Mills, NBME

Global Considerations

Maria Elena Oliveri, Buros Center for Testing-UNL

Security Guidelines for Technology-Based Assessments

David Foster, Caveon Test Security

Guidelines for Reporting Results from Technology-Based Assessments

April Zenisky, University of Massachusetts Amherst

Implications for the Technology-Based Guidelines for Educational Assessment

Chad W. Buckendahl, ACS Ventures, LLC

Discussant:

John Weiner, PSI Services, LLC

(SIGIMIE Session) Testing Time: The Push and Pull in High-Stakes State Accountability Assessments

11:00 to 12:30 pm - Organized Discussion

Due to the increased rigor of statewide curriculum standards, and to the public outcry against "over-testing" students in public schools, "testing time" has become a regular conversation in state education agencies and amongst the measurement community. New content standards call for students to demonstrate mastery of the deep, rich skills intended to be taught in schools. This development has caused state education agencies, assessment consortia, and the measurement community to build tests aligned to these expectations. Unfortunately, though not entirely unexpectedly, these tests greatly increased the amount of time students spent on their annual state accountability exams, which has put many state agencies in the difficult position of being required to reduce testing time, maintain the initial construct the assessments were intended to measure, support reporting structures that educators find useful, allow for comparability to support the continuation of accountability systems, and become entrenched in state politics like never before. In this session, we will address the issue of testing time head on by having a presentation from a state assessment director—followed by a blue-ribbon panel to discuss these issues. Audience discussion will also be facilitated.

Session Organizer:

Vince Verges, Florida Department of Education

Presenters:

Andrew J. Middlestead, Michigan Department of Education
Susan Brookhart, Duquesne University
Kristen Huff, Curriculum Associates
Mark Reckase, Psychometric Solutions
Joyce Zurkowski

Advancing Assessment in Medical Education
11:00 to 12:30 pm - Coordinated Paper Session

This coordinated session provides an overview of advances made in medical education assessment that (1) improve upon test development and test score reporting, (2) use process data to obtain new insight into examinee response processes, and (3) introduce new methods of assessing communication constructs or assessing new constructs in communication skills. More specifically, topics in this session will report on: development of a score reporting dashboard to meet the needs of various stakeholders; the utility of applying a bipartite graphs to an examination to address missing scores ; the relationship between item characteristics and answer changes using the integrated hierarchical-speed-accuracy and answer change model; the use of an exam delivery feature that allows examinees "strike-out" response options they believe to be incorrect; the utility of automatic speech recognition to improve the assessment of clinical skills; and a program of construct validation for patient-centered written communication. Together, this research showcases new assessment capabilities for producing high quality, high-stakes medical education assessments.

Session Organizer:
Christopher Runyon, NBME

Chair:
Su Somay, NBME

Participants:
The Relationship Between Answer Changes and Item Characteristics on a High-Stakes Examination
Aaron Myers, University of Arkansas; Irina Grabovsky, National Board of Medical Examiners
Examining the Use of Strikeouts on a Computerized High-Stakes Examination
Ravi Pandian, National Board of Medical Examiners; Polina Harik, National Board of Medical Examiners
Advances in Reporting Results on Medical Education Assessments
Francis O'Donnell, National Board of Medical Examiners; Thai Quang Ong, National Board of Medical Examiners; Rich Feinberg, National Board of Medical Examiners
Assessing Patient-Centered Written Communication
Amanda Clauser, National Board of Medical Examiners; Ann King, NBME
Leveraging Automatic Speech Recognition to Advance Assessment of Clinical Skills for Medical Licensure
Su Somay, NBME; Victoria Yaneva, NBME; Melissa Margolis, National Board of Medical Examiners
Development of a Performance Recommendation System Using Bipartite Graphs
Janet Mee, NBME; Amanda Clauser, NBME, & Christopher Yang, Drexel University

Discussant:
Thanos Patelis, Fordham University, Teachers College, University of Kansas

(SIGIMIE Session) Challenges and opportunities in delivering virtual oral and OSCE examinations

2:00 to 3:30 pm - Organized Discussion

In this first session ever sponsored by the NCME Certification and Licensure SIGIMIE, four panelists from four certification and licensure organizations are gathered for a discussion on the practical challenges and research opportunities provided in moving face-to-face examinations to an online, virtual format. Oral and Objective Structured Clinical Examinations (OSCEs) are common components of licensure and certification requirements. The COVID-19 pandemic created obvious safety concerns for administering such examinations, forcing testing organizations to quickly and creatively alter the nature of these exam modalities while striving to maintain reliability and validity. Four panelists from the certification and licensure industry with direct experience addressing these challenges will assemble a panel to discuss both the practical (form publication, content security, technology platforms) and psychometric (score comparably, rater consistency, construct stability) challenges of rapidly repositioning the delivery of high-stakes, operational examinations. Our experienced discussant will direct the conversation to ensure a centered grounding using Kane's validity framework on how modality shifts may support or challenge existing score use arguments.

Session Organizers:

Mikaela Raddatz, American Board of Physical Medicine and Rehabilitation
Jonathan Rubright, National Board of Medical Examiners

Presenters:

Amanda Clauser, National Board of Medical Examiners
Brian J Hess, College of Family Physicians of Canada
Andrew Jones, American Board of Surgery
Sarah Schnabel, American Board of Ophthalmology

Creating Coherence: Integrating Principled Assessment Design, PLDs, and Standard Setting
2:00 to 3:30 pm - Coordinated Paper Session

Performance Level Descriptors (PLDs) are intended to provide detailed descriptions of what knowledge and skills students at each performance level (e.g., basic, proficient, advanced) possess. Under a Principled Assessment Design (PAD) framework, PLDs should be developed not just prior to standard setting, but prior to writing items. Test developers should start out with a hypothesized set of PLDs, write items which elicit evidence for these descriptors, and then evaluate and iteratively revise both items and the PLDs based on empirical data obtained during field testing. This approach works well for new assessment programs which are starting from scratch; however, this does not mean that assessment programs which are already "in flight" cannot benefit from incorporating PAD approaches into their test development process. This session will include presentations of three papers which discuss efforts to implement PAD practices for PLD development, item development, and standard setting for the i-Ready Diagnostic assessment program. Presenters in this session will describe the unique challenges with implementing PAD for an assessment program with an existing item bank and discuss proposed solutions which have been investigated to address them—both within the PLD development process and within the cut score determination process.

Session Organizer:
Laurie Davis, Curriculum Associates

Participants:
Practical and Principled PLD Design: Balancing and Building Evidence and Expertise
Amanda Brice, Curriculum Associates; Daniel Mix, Curriculum Associates
Analytic Methods and Tools Supporting PLD, Item alignment, and Cut Score Coherence
Daniel Lewis, Creative Measurement Solutions LLC
How to Win at Standard Setting with Imperfect Data
Matthew N. Gaertner, WestEd; Sonya Powers, WestEd

Discussant:
Ellen Forte, edCount, LLC

Graduate Student Reception
8:00 pm
Hosted by the Graduate Student Issues Committee

Join the Graduate Student Issues Committee (GSIC) at a virtual get together with other NCME graduate students!

(Invited Session) Using Educational Assessments to Educate: Opportunities for Leveraging the "Power" of Assessment

9:00 to 10:30 am - Coordinated Paper Session

Many psychometricians and educational researchers talk about the need to integrate teaching and assessment, but examples of such integration are hard to find. Thus, the current state of affairs is a lot of discussion regarding how the power of educational assessment can help improve educational outcomes for all students, with very few examples of how that can be done. In this session for the 2021 conference we bring together an all-star team of education researchers and measurement experts to illustrate how we can improve educational assessments to support the education of all students, particularly those from traditionally underserved populations. The topics discussed in these presentations aim to move the measurement beyond 20th century ways of conceptualizing the constructs that are important to measure, the ways in which we measure them, and our validation frameworks to support the inferences and uses we prioritize. The presentations, discussant remarks, and audience interaction will help practitioners to better develop and validate educational assessments that will benefit students from all walks of life.

Session Organizer:

Stephen Sireci, University of Massachusetts Amherst

Chair:

Eva Baker, UCLA

Participants:

The Economic Consequences of Ignoring Testing Consequences

Fernanda Gandara, University of Massachusetts Amherst; Stephen Sireci, University of Massachusetts Amherst

Learner-centered Assessments Through Dynamic Pedagogy

Eleanor Armour-Thomas, Queens College of the City University of New York

Creating Opportunities to Learn Through Catalytic Assessments

Cynthia McCallister, New York University

Proximal, Diagnostic, and Formative Assessments that Learners and Teachers Can Use!

Madhabi Chatterji, Teachers College, Columbia University

The Influence of Automaticity on Reading and Mathematics Achievement: A Multicultural Analysis of ECLS-K9

Ernest Washington, University of Massachusetts Amherst; Ted Daisher, University of Massachusetts Amherst; Stephen Sireci, University of Massachusetts Amherst

Discussant:

Edmund Gordon, John M. Musser Professor of Psychology, Emeritus - Yale University / Richard March Hoe Professor

From CAT to Smart Learning – Urgent Research During the Pandemic

9:00 to 10:30 am - Coordinated Paper Session

The theory and methods of Computerized Adaptive Testing (CAT) have been well advanced during the last 40 years. The rapid developments in technology have made large-scale CAT implementations easier than ever before. However, CAT methods have not been well acknowledged or even known by many researchers in Artificial Intelligence (AI). Today we would like to highlight our discussion whether CAT can help AI in educational research, in particular, how to better design "deep learning" and "neural network" to archive many attractive missions, such as smart testing and smart learning. We will show how CAT can be utilized to build a tailored assessment for each individual in the big data era. Our goal is to build many reliable, and also affordable, web-based diagnostic tools for schools to automatically classify students' mastery levels for any given set of cognitive skills that students need to master. With the COVID-19 pandemic, many schools have canceled in-person classes and moved to online instructions, which has created enormous challenges for both teachers and students. This coordinated session will clearly show that the new technologies can be immediately employed to support individualized learning on a mass scale whether in-class, online or hybrid.

Session Organizer:
Hua-Hua Chang, Purdue University

Chair:
Wen Zeng, Cambium Assessment, Inc.

Participants:
Urgent CD-CAT Research During The COVID-19 Pandemic
Hua-Hua Chang, Purdue University; Wen Zeng, Cambium Assessment, Inc.
AI and Machine Learning in Psychometrics? Old News
Nathan Thompson, Assessment Systems Corporation
On-The-Fly Parameter Estimation Based on Item Response Theory in Adaptive Learning Systems
Chun Wang, University of Washington; Shengyu Jiang
Automated Attribute Hierarchy Detection with Application to Adaptive Learning
Yinghan Chen, University of Nevada, Reno; Shiyu Wang, University of Georgia
Understanding Interactive Items' Characteristics by Deep Learning-based Process Data Analysis
Susu Zhang, University of Illinois at Urbana-Champaign

Discussant:
Hua-Hua Chang, Purdue University

(Invited Session) Artificial Intelligence, Machine Learning, and the Future of Assessment

9:00 to 10:30 am - Organized Discussion

The field of data science has grown considerably over the last several years. Increased computing capabilities have led to advances in artificial intelligence and machine learning which allow for the automation of tasks which previously had to be curated by humans. These advances promise to revolutionize how we use and interpret data. Some of these applications have already become part of measurement landscape, such as automated scoring of performance tasks, and automatic item generation. Other applications are beginning to enter the measurement field at the periphery, often in unregulated ways. The purpose of this moderated panel session is to discuss the ramifications of these advances for the future of assessment. Panel members will discuss several of the statistical models and estimation algorithms being employed in AI, as well as related implications for the interpretability of results, potential bias, ethical concerns, and data management. The discussion will connect these topics to this year's NCME theme by focusing on the research that supports these methods, and how they might best be put into practice.

Session Organizer:

Billy Skorupski, Questar Assessment

Presenters:

Billy Skorupski, Questar Assessment
Susan Lottridge, Cambium Assessment
David Williamson, College Board
Victoria Yaneva, National Board of Medical Examiners
John Whitmer, Federation of American Scientists
Victoria Vassileva, Arthur AI

High Definition Detection of Testing Misconduct

9:00 to 10:30 am - Coordinated Paper Session

My son asked me whether to choose SD or HD after I entered my credit card number for his movie rental. “HD, because the picture is sharper and clearer”, I told him without hesitation. The goal of this symposium is to offer new and improved tools to paint a high definition picture to help make measurement matter more. Measurement does not likely matter if the integrity of test scores is questionable. Who would use test scores that are questionable, invalid, or unfair? Misconduct in educational testing, however, do occur from time to time, resulting in an unfair advantage for some test takers. Thus, it is crucial to run test security analyses to detect aberrant response behaviors, unusual response similarity, or abnormal item performance. Much effort has been made to detect these situations under different settings, including the new and improved methods proposed in this symposium. Making detection more accurate is the same goal shared by five groups of presenters from different universities and testing companies. They will share their findings on ranking response time models, reinforcing the sequential procedure, recommending a machine learning approach, refining raw data, and redefining the longest identical string.

Session Organizer & Chair:

Zhongmin Cui, CFA Institute

Participants:

Impact of RT Model Selection in Detecting Aberrant Test-taking Behaviors

Huijuan Meng, AWS

Hybrid Threshold-Sequential Procedures for the CAT Security

Chansoon Lee, Liberty University; Hong Qian, National Council of State Boards of Nursing

A Weak Supervised Learning Approach for Detecting Item Preknowledge in CAT

Yiqin Pan, University of Wisconsin-Madison; Edison M. Choe, Graduate Management Admission Council

Improving Test Security Analysis Through Noise Removing

Mingjia Ma, University of Iowa; Zhongmin Cui, CFA Institute

Making the Longest Identical String Longer

Zhongmin Cui, CFA Institute

Discussant:

James Wollack, University of Wisconsin

Issues in Item and Test Design

9:00 to 10:30 am - Paper Session

Chair:

Yong He, ACT

Participants:

Comparing Unfolding Medical Case Studies to Their Individual Single-Item Counterparts

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

Unfolding medical case studies are well-suited for measuring decision-making skills (e.g., clinical judgment). As case studies unfold, newly introduced information shifts focus across different medical topics. This counteracts inter-item dependencies, potentially reducing it. Unfolding case studies are compared to discrete item counterparts to explore the utility of dynamic item sets.

Moving from Writing Items to Designing Items to Provide Better Assessment Information

Weeraphat Suksiri, University of California, Berkeley; Linda Morell; Mark Wilson, University of California, Berkeley

Item design is a special observation ensuring that intended information is elicited by items. Item designing is foundationally and practically different from item writing. This study shows how intended information of a complex skill assessment is obtained from item designing procedure and how the procedure supports validity of score interpretation.

Crafting an Unfolding Model for Measuring Pedagogical Content Knowledge

Jiwon Nam-Speers, University of Baltimore; Nan Sook Yu, Chonnam National University; Hyejin Shim, University of Missouri

The responses to affective testing items could be more congruent with unfolding mechanism, while those to cognitive items follow cumulative mechanism. This study aimed to compare Rasch and 2PL IRT models and unfolding model. As result, both 2 PL and unfolding models had good model-fit, while Rasch Model did not.

Literature Review of Situational Judgment Tests for High-Stakes Selection

Ted Daisher, University of Massachusetts Amherst; Amanda Clauser, National Board of Medical Examiners

This review summarizes literature relevant to the use of situational judgment tests (SJT) for high-stakes selection and medical training. Studies were gathered through a database search and snowballing. The review focuses on developing SJTs and gathering validity evidence, illustrating with examples from SJTs used for selection in medical training.

The End of One-Size-Fits-All Testing: Personalizing Test Schedules with Recommender Systems

Okan Bulut, University of Alberta; Damien Cormier, University of Alberta; Jinnie Shin

This study aims to demonstrate the utility of recommender systems in generating automated and personalized test schedules for computer-based formative assessments. Using a large sample of students, we show that the recommender system can reduce the number of test administrations significantly by generating a personalized testing schedule for each student.

Discussant:

Sarah Linnea Toton, Caveon Test Security

The Resurgence of Interim Assessment—Bringing Teaching and Testing Back Together

9:00 to 10:30 am - Organized Discussion

Following the passage of No Child Left Behind in 2001, K-12 testing experienced a massive expansion as states sought to leverage end-of-year summative tests as tools for accountability purposes. A public and grass roots backlash of teachers and parents against over-testing has fueled the opt-out movement that has swept the country in recent years. One of the complaints is that summative tests generally do not provide the type of data that teachers find valuable in terms of informing instructional decisions. Summative testing has an important role in educational policy, but the results are not sufficiently granular and come too late in the school year to provide actionable feedback to improve instruction. Interim assessments are typically given multiple times a year and tend to be perceived as lower stakes than summative assessments. As such, they can be designed to support teaching and learning throughout the year. Interim assessments can be used by educators to evaluate student progress, to create instructional groupings, to support decisions about whole class instruction, and to support personalized learning. In this session, presenters will discuss how interim assessments can complement classroom, formative, and summative assessments to provide teachers actionable information to drive learning.

Session Organizer:

Laurie Davis, Curriculum Associates

Presenters:

Laurie Davis, Curriculum Associates
John Denbleyker, Houghton Mifflin Harcourt
Katie McClarty, Renaissance
Karen Barton, NWEA
Michelle Derbenwick Barrett, Edmentum

Innovations in Response Time Models

9:00 to 10:30 am - Paper Session

Chair:

Na Liu

Participants:

Investigating Speed-Accuracy Relationships Using Latent Profile Analysis

Tanesia Beverly, Law School Admission Council; Eric Loken, University of Connecticut; Alexander Weissman, Law School Admission Council

The relationship between response time and accuracy was investigated using responses from a large-scale test. Using mixture modeling, we identified multiple patterns for the speed-accuracy trade-off. These findings have implications for joint modeling, as a single covariance structure may not capture the relationship's dynamics.

An Investigation of Differential Item Timing Functioning in Digitally Based Assessments

Young Yee Kim, American Institutes for Research; Xiaying Zheng, American Institutes for Research; Xiaoying Feng, American Institutes for Research; Nixi Wang; Markus Broer, American Institutes for Research

Research has found differences in item response time between groups, suggesting potential differential item timing functioning (DITF). The purpose of this study is investigating feasibility of DITF analysis using NAEP data, by examining if there are differences in response time between two groups, conditioned on their latent ability and speed.

A Response-Time-based CUSUM detection of aberrant testing behaviors

Yang Du; Onur Demirkaya, University of Illinois at Urbana-Champaign; Justin L. Kern, University of Illinois at Urbana-Champaign

Responding with item preknowledge can greatly undermine test validity. To detect these aberrant behaviors, we propose four response-time-based statistics relying on the CUSUM procedure. Two simulations and an empirical study were conducted. Our results show that the T2 and T4 showed higher power than the other two statistics.

Using Response Time Modeling to Detect Speeded Examinees with Missing Responses

Xiaying Zheng, American Institutes for Research; Tong Wu; Young Yee Kim, American Institutes for Research; Fusun Sahin, American Institutes for Research

Conventionally test speededness is identified using missing responses. This study uses joint modeling of response time and response data to identify speeded examinees by estimating their expected response time under no time constraint. This research will contribute to the development of evidence-based methods to identify speeded examinees, and ultimately, tests.

Detecting Examinees' Item Pre-knowledge: A New Lognormal Response Time Model

Murat Kasli, University of Miami; Cengiz Zopluoglu, University of Miami; Nooree Huh, ACT

A novel deterministic gated lognormal response time (DG-LNRT-N) model is proposed to detect students with item pre-knowledge. The efficacy of the new model will be demonstrated through simulation by manipulating sample sizes, a number of items, percentages of compromised items, and examinees with item pre-knowledge, based on real test data.

Discussant:

Michelle Boyer, Center for Assessment

Wednesday Coffee Chat Sessions

10:35 to 11:00 am

Join your NCME colleagues for a unique opportunity to share ideas, questions, and thoughts about current topics in our industry.

1. **Coffee Chat: Opportunities for Assessments to Serve Education**
   Hosted by Steve Sireci, University of Massachusetts Amherst

2. **Coffee Chat: Public Perceptions of our Work in Measurement and the Road Ahead**
   Hosted by Emily Shaw, College Board

3. **Coffee Chat: Ethics and Assessment Data**
   Hosted by Tiago A. Caliço, American Institutes for Research

4. **Coffee Chat: Assessing Special Student Populations**
   Hosted by Edynn Sato, Sato Education Consulting LLC

5. **Coffee Chat: Non-Cognitive Competencies in Credentialing**
   Hosted by Joanne Kane, National Conference of Bar Examiners

6. **Coffee Chat: Interim Assessment: Meeting in the Middle?**
   Hosted by Laurie Davis, Curriculum Associates

7. **Coffee Chat: Chill with a little Chat**
   Give your mind a little rest in this 25-minute break. We'll say hello and settle in for the first 5 mins, have a guided meditation for 15 mins, and use the last 5 mins for a little chat before we transition to the next session.
   Hosted by: Rosemary Reshetar, National Conference of Bar Examiners

(CODIT Feature Session) Black Lives Matter in Educational Measurement

11:15 to 12:45 pm - Organized Discussion

This organized discussion panel calls for a unified and deliberate commitment to anti-racist assessment and measurement in the midst of a pandemic that is disproportionately impacting Black communities and killing Black Americans. Indeed, as we witness anti-Black racist violence in this country through the murders of Black citizens such as George Floyd and Breonna Taylor, this anti-Black violence is mirrored in American K-16 classrooms daily. As the current AERA president Shaun Harper notes: "Evidence from multiple sources across numerous academic disciplines and fields consistently highlights systems that cyclically disadvantage Black people." Indeed, schooling is one of those systems. It is within this context that educators of educational measurement must do our work. Simply saying "Black Lives Matters" is not enough. As an organization, and as educators, we must unite in meaningful ways to disrupt racist practices and policies - especially those perpetuated through educational assessment. In this current sociopolitical context, the questions become - How has Black Lives Mattered in the context of measurement education? - How has Black Lives Mattered in our research, scholarship, teaching, disciplinary discourses, graduate programs, professional organizations, and publications? - How can we, as a discipline, contribute to the political freedom of Black peoples?

Session Organizers:

Jennifer Randall, University of Massachusetts
Kristen Huff, Curriculum Associates

Presenters:

Kyndra Middleton, Howard University
Mya Poe, Northeastern University
Kerrita Mayfield, Amherst Public Schools

(SIGIMIE Session) Navvy Education: Building and Implementing a Statewide Diagnostic Assessment System

11:15 to 12:45 pm - Organized Discussion

The Navvy assessment system ("Navvy") is an online, classroom-embedded diagnostic assessment system created by Dr. Laine Bradshaw, associate professor at the University of Georgia, and is currently operating in districts across the state of Georgia. Navvy is being piloted in Georgia under the federal Innovative Assessment Demonstration Authority (IADA). Navvy is designed to give detailed feedback about students' understandings of specific standards, or learning targets, by using diagnostic measurement models. Assessments can be administered as needed throughout the year and results are provided in real-time, so the diagnostic feedback can be used immediately by teachers to tailor instruction towards students' individualized needs. This session will provide the theory of action underlying the design of Navvy, introduce the fundamentals of the system, overview the implementation of Navvy in collaboration with school districts and key stakeholders, and discuss some of the opportunities as well as hurdles of implementation. Presenters include the Navvy assessment development and psychometric team members and leaders of districts in the Navvy consortium. A moderated panel of diagnostic assessment researchers and practitioners will debate and discuss the past, present, and future of diagnostic measurement, the Navvy assessment system, other diagnostic assessment systems, and outstanding challenges in the field of diagnostic measurement. Audience questions will be encouraged.

Chair:

Benjamin R. Shear, University of Colorado Boulder

Presenters:

Laine Bradshaw, University of Georgia
Eric Arena, Putnam County School District
Amanda Miller, Scintilla Charter Academy
Brooke Knight, Scintilla Charter Academy
Jennifer White, Floyd County Schools
John Parker, Floyd County Schools
Matthew James Madison, University of Georgia
Andre Rupp, Mindful Measurement

Advancing Digital Instruction and Assessment with Natural Language Processing & Learning Analytics

11:15 to 12:45 pm - Coordinated Paper Session

Digital instruction and assessment research has gained traction in recent years. COVID-19 has significantly impacted educational practice, pushing educators and their students to necessarily embrace digital education. With greater use of educational technology, increased digitally-captured student data will be available for research. New knowledge from this research can be further leveraged to advance digital instruction and assessment practice. Presentations in this symposium demonstrate how natural language processing (NLP) technology and learning analytics can support large-scale research of educational data and inform educational practice. The presentations will address these research questions: (1) How do NLP features from student writing on timed writing assessments compare to coursework writing?; (2) How can digitally-captured process and product data from writing tasks be leveraged for formative feedback?; (3) How can we use machine learning to discover user behavior patterns from process and product data from a digital writing app, and demonstrate relationships between user behaviors and writing quality?; and, (4) How can we measure learning in collaborative learning environments, educational simulations, and intelligent tutoring systems? Using the domains of writing and collaboration, the presentations will demonstrate how NLP and learning analytics research contribute to the advancement of digital instruction and assessment practice.

Session Organizer:

Mo Zhang, Educational Testing Service

Chairs:

Mo Zhang, Educational Testing Service
Jill Burstein, Educational Testing Service

Participants:

Are Standardized Writing Assessments Representative of Students' Writing?

Daniel McCaffrey, Educational Testing Service; Mo Zhang, Educational Testing Service; Jill Burstein, Educational Testing Service; Beata Beigman Klebanov, Educational Testing Service; Steven Holtzman, Educational Testing Service; Norbert Elliot, New Jersey Institute of Technology

Uncovering Patterns of Use in the Writing Mentor® App through Cluster Analysis

Mengxiao Zhu, Educational Testing Service; Jiangang Hao, Educational Testing Service; Jill Burstein, Educational Testing Service; Oren Livne, Educational Testing Service

Measuring Collaborative Learning: Design, Data, and Methods

Alina A. von Davier, ACT

Discussant:

Danielle McNamara, Arizona State University

Psychometric Modeling of Data Based on a Table of Specifications

11:15 to 12:45 pm - Coordinated Paper Session

This session considers different approaches to modeling examinee data based on a table of specifications (TOS) framework. The first paper provides a snapshot of how the TOS for the latest edition of the Medical College Admissions Test was developed and highlights the important role it plays throughout the exam lifecycle. The second paper introduces an extension of Multivariate Generalizability Theory (MGT), called "Extended" MGT (XMGT). XMGT is more comprehensive in the sense that it can handle more complex TOS designs compared to MGT, provides estimates of conditional standard errors of measurement for raw and scale scores, and provides an indicator of model-data fit. The third paper illustrates how XMGT can be used to model a TOS for a multi-faceted exam and contrasts its use to that of univariate GT. Using the same data as the previous paper, the fourth paper models the TOS using item response theory (IRT) employing both unidimensional and multidimensional models. Results from the third and fourth papers are compared. In the last segment of the session, an expert in psychometrics discusses each of the four aforementioned papers and presentations.

Session Organizer:

Jaime Malatesta, Graduate Management Admission Council

Participants:

The Role of Test Specifications in the Exam Lifecycle

Marc Kroopnick, Association of American Medical Colleges; Ying Jin, Association of American Medical Colleges; Cynthia Searcy

Extended Multivariate Generalizability Theory

Robert Brennan, University of Iowa

Using Extended Multivariate Generalizability Theory to Model a Table of Specifications

Jaime Malatesta, Graduate Management Admission Council; Robert Brennan, University of Iowa; Won-Chan Lee, University of Iowa

IRT Approaches to Modeling a Table of Specifications

Stella Kim, University of North Carolina at Charlotte; Won-Chan Lee, University of Iowa

Discussant:

Michael Kolen, University of Iowa

Validity, Psychometric Properties, and Accessibility of Innovative Item Types in K-12 Assessments

11:15 to 12:45 pm - Coordinated Paper Session

With extensive applications of advanced technology in online testing, innovative items have been widely explored and implemented in online testing. These new item types promote substantial changes in item structure, response style, and offer interactive activities in K-12 assessment. Innovative item types intentionally measure higher level of cognitive complexity and are usually scored with partial-credit models. The psychometric properties of innovative items, however, are rarely studied with empirical evidence and reported in measurement literature. The current session incorporates three empirical studies on innovative item types. Using simulated and operational data from state assessments, validity, psychometric properties, and scoring of a variety of innovative item types are investigated in online testing. Advantages and practical issues of using those new item types to measure student performance, especially for students with disabilities, are discussed for the improvement in item development, scoring, accessibility, and the technical quality of assessments. A highly regarded discussant in the areas of measurement and K-12 assessments will provide comments on the three studies, strengths and weaknesses and discuss innovative item types and their implementations in online testing. Abstract: 178 words

Session Organizer:

Liru Zhang, Assessment Consulting Services

Chair:

Liru Zhang, Assessment Consulting Services

Participants:

Validity and Psychometric Properties of Integrated Item Cluster in Science Assessments

Liru Zhang, Assessment Consulting Services; Shudong Wang, NWEA

An Investigation of Efficiency and Validity Evidence in Scoring Innovative Items

Shudong Wang, NWEA

Evaluating Partial Credit Scoring of Multiple Select Items

Mark Hansen, UCLA; Eric Setoguchi, National University; Matthew Schulz, Smarter Balanced Assessment Consortia

Discussant:

Richard Patz, University of California, Berkeley

Application of Fit Statistics

11:15 to 12:45 pm – Paper Session

Chair:

Yi-Fang Wu, ACT

Participants:

Person Fit z-statistics for Rasch Testlet Model

Zhongtian Lin, Cambium Assessment, Inc; Tao Jiang, Cambium Assessment, Inc; Frank Rijmen, Cambium Assessment, Inc

Person fit z-statistics lz and lz* were extended for the Rasch testlet model when the examinee ability is estimated by a marginal maximum likelihood estimator. Simulation results showed that lz* has close-to-nominal Type I error rates, and satisfactory power for detecting aberrant responses.

Impact of Item Misfit on Group Score Reporting in Large-Scale Assessments

Seang-Hwane Joo, Educational Testing Service; Usama Ali, Educational Testing Service; Frederic Robin, ETS

We investigate the potential impact of various types of item misfit via simulated data that mimics the empirical large-scale assessments. A real-data-based simulation study is conducted to manipulate the type and magnitude of misfit and evaluate their impact of item misfit on the group-level score and scale comparability.

Application of Machine Learning to Detection of IRT Item Misfit

John Donoghue, Educational Testing Service; Bingchen Liu

Machine learning models were used to predict IRT item fit in a large-scale assessment program. Using IRT calibrations from several years, multiple ML models predicted whether items demonstrated enough misfit to require treatment. Administration and item meta-data, and four item-fit measures were features. K-fold cross-validation and holdout-sample performance were criteria.

Person-Fit Statistics for A Joint Testlet Model for Accuracy and Speed

Wei Xu, National Council of State Boards of Nursing; William J Muntean, National Council of State Boards of Nursing

Aberrant test-taking behaviors can significantly diminish validity of test results. In this study, the performance of the proposed joint testlet was compared with that of a joint baseline model that did not account for testlet effect using simulation data. Person-fit statistics were utilized to evaluate aberrant behaviors.

IRT Item Fit Statistics Based on Item Response Residuals

Scott Monroe, University of Massachusetts Amherst

This study proposes several IRT item fit statistics based on item response residuals. The statistics are developed using the generalized residuals framework (Haberman & Sinharay, 2013), and thus have known asymptotic reference distributions. A simulation study shows that two of these statistics perform well for the 2PL model.

Discussant:

Dubravka Svetina Valdivia, Indiana University

Application of Response Time Models

11:15 to 12:45 pm - Paper Session

Chair:

Yong Luo

Participants:

Is it Better to Imperfectly Classify Rapid Guessing Than to Ignore it?

Joseph A. Rios, University of Minnesota

A failure to accurately classify rapid guessing via the use of response latencies leads to the inclusion of distortive and psychometrically uninformative information in parameter estimates. To better understand this degree of bias, a simulation study was conducted in which misclassification types and rates of rapid guessing were manipulated.

Investigating rapid responses incorporating multiple-choice alternatives using IRT Nominal Model

Mingjia Ma, University of Iowa; Stephen B. Dunbar, University of Iowa; Catherine Welch, University of Iowa; Ylibo Wang

Rapid-guessing behavior has been investigated with the availability of response time data and interpreted as a result of low-motivation from test-takers where they randomly pick an alternative to answer the question. Results of this paper show that those rapid responses might be resulted from highly distractive alternatives.

Joint Modeling of Responses, Response Time, and Answer Changes for Cognitive Diagnosis

Hong Jiao, University of Maryland; Yishan Ding, University of Maryland, College Park; Chengbin Yin, University of Maryland, College Park; Shudong Wang, NWEA

Built upon previous studies on joint modeling of responses and response time for cognitive diagnosis, this study adds another data type, answer change data in the joint modeling for cognitive diagnosis. Different approaches to model answer changes are explored. Different joint models are proposed and model parameter estimation is investigated.

Using Response Times for Modeling Careless Responding and Attentive Response Styles

Esther Ulitzsch, Leibniz Institute for Science and Mathematics Education; Lale Khorramdel, Boston College; Ulf Kroehne; Steffi Pohl, Freie Universitat Berlin; Matthias Von Davier, NBME

A response time based mixture model for questionnaire data is presented that identifies careless respondents and accounts for extreme and midpoint response styles in attentive responding. Its utility for more valid conclusions concerning measured traits and insights into response behavior is illustrated using empirical and simulated data.

Detecting Careless Responses: A New Method Utilizing Response Time and Accuracy

Seyma Nur Yildirim-Erbasli, University of Alberta; Guher Gorgun, University of Alberta; Okan Bulut, University of Alberta

This study proposes a new method for handling rapid-guessing behavior on test items by taking response accuracy into account. We compare the performance of the method with those of the existing method using real data from a large-scale formative assessment. Results indicate that the new method outperformed the existing method.

Discussant:

Michelle Derbenwick Barrett, Edmentum

Electronic Board Session #1

1:00 to 2:00 pm

Participants:

A Semi-Confirmatory Latent Dirichlet Allocation Topic Model

Jordan M. Wheeler, University of Georgia; Hye-Jeong Choi, University of Georgia; Jiawei Xiong, University of Georgia; Allan Cohen, University of Georgia; Juyeong Lee, University of Georgia

Topic models are statistical models that freely estimate the underlying semantic structure of a set of textual data. In this study, we propose a new semi-confirmatory Latent Dirichlet Allocation topic model in which we fix a subset of the topics a priori. Its utility is demonstrated using an empirical example.

Prediction of Item Difficulty Using Natural Language Processing with Topic Modeling

Minju Hong, University of Georgia; Yanyan Fu, GMAC; Kyung (Chris) Han, Graduate Management Admission Council

This study applied topic modeling to predicting item difficulty of verbal reasoning items in a large-scale assessment. The topics of the items extracted from latent Dirichlet allocation model were used as predictors in a multiple regression model. The results showed that the topics were significant predictors of item difficulty.

A Modified S-X^2 Statistic Accounting for Sample Size

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

Chi-squared statistics are known to be sensitive to sample sizes such that large sample sizes can make small differences appear statistically significant. This study suggests a modified S-X^2 statistic that accounts for sample-size issues and aims to provide practical implications about optimum sample sizes that can yield appropriate S-X^2 results.

Analyzing Technology Enhanced Items Using Innovative Approaches

Ji Zeng, Michigan Department of Education; Warren Li, University of Michigan-Ann Arbor; Tamara Smolek, Michigan Department of Education

Technology enhanced (TE) items have become more prevalent on state summative assessments. However, TE raw responses are “text” like string variables which are difficult to recode. We computed TF-IDF for a drop-down TE item to show how we may simplify the recoding step to facilitate TE item analysis.

An Examination of NCDIF Index for Detecting Item Parameter Drift (IPD)

Juan Chen, National Conference of Bar Examiners; Won-Chan Lee, University of Iowa; Mark R Connally, National Conference of Bar Examiners; Mark Albanese, National Conference of Bar Examiners

This simulation investigated Type I error and power of NCDIF in detecting IPD using three IPR procedures when sample size, features of IPD, weights, and examinee abilities were manipulated. Results indicate that modified IPR procedures performed better for unequal-sample conditions; features of IPD and examinee ability distributions impacted IPD detection.

Application of Mixture Explanatory Item Response Models to Explore Response Process Validity

Clifford Erhardt Hauenstein; Eunbee Kim, Georgia Institute of Technology; Susan Embretson, Georgia Institute of Technology

Response process validity of a psychometric diagnostic tool is explored through mixture versions of explanatory item response models. Fluid reasoning/rule derivation items are reconceptualized as Boolean expressions to identify the relationship between item structure and difficulty. This relationship is then used to derive latent class structure with several competing models.

Application of Multidimensional Mixture IRT and supervised-LDA for DIF in mixed-format test

Juyeon Lee; Jiawei Xiong, University of Georgia; Jordan M. Wheeler, University of Georgia; Hye-Jeong Choi, University of Georgia; Allan Cohen, University of Georgia

A multidimensional mixture IRT model is proposed to detect DIF in mixed-format tests. A supervised LDA model is incorporated to assist in interpretation of latent classes in constructed response item.

Application of the Multidimensional Latent Regression Model to Test Scoring

Preston Botter, UCLA CRESST

In the context of language testing, we apply the multidimensional latent regression model to obtain improved domain estimates of proficiency.

Developing a Survey to Assess At-Home Spatial Reasoning

Anthony Sparks; Sarah Wellberg, University of Colorado Boulder; Josh Geller, University of Oregon; Jennifer McMurrer, Southern Methodist University; Cassandra Hatfield, Southern Methodist University; Leanne Ketterlin Geller, Southern Methodist University

The current study is focused on the development and validation efforts of a home environment survey for the purpose of measuring children's spatial reasoning from the perspective of a parent/guardian. We discuss the iterative development cycle and evidence collected to refine the instrument.

Employable Skills Self-Efficacy Survey: A Validation Study

Amanda Rose Dumoulin, Kwantlen Polytechnic University; Shayna Rusticus, Kwantlen Polytechnic University

We conducted a validation study on the Employable Skills Self-Efficacy Survey (Ciarocco & Strohmetz, 2018) with 170 undergraduates. Results did not support the proposed factor structure. We found evidence of convergent, but not discriminant validity. Revisions are necessary prior to using scale to measure employable skills self-efficacy of psychology students.

Evaluating Different Scoring Methods for Multiple Response Items Providing Partial Credit

Joseph Betts, National Council of State Boards of Nursing; William J Muntean, National Council of State Boards of Nursing; Doyoung Kim, National Council of State Boards of Nursing; Shu-chuan Kao, NCSBN

The multiple response structure can underlie several different TEI response methods, e.g. highlighting, drag-and-drop, etc. This presentation will provide the results of using several polytomous scoring methods. Each scoring method will be discussed in-depth and results applicable to many operational programs.

Examination of Reliability, Sample Size, and Test Length on DIF Detection Methodology

Jinmin Chung, University of Iowa; Ye Ma, University of Iowa; Terry Ackerman, University of Iowa

This study examines how reliability, sample size, and test length affect the performance of various DIF methods, which are Mantel-Haenszel, SIBTEST, Raju's DFIT, and Lord's Chi-square. This study raises awareness of test reliability when conducting DIF analysis. The conclusion can help understand DIF results based on test's psychometric properties.

Examining Educator Actions in Response to Student Rapid Guessing Alerts

Audra Kosh, Edmentum

Using item response time, computer-based tests often alert educators when a student may have demonstrated non-effortful behavior. This study examines the actions educators take in an online environment following such an alert and the resulting effects on students.

Exploring Various Approaches to Transforming a Linear Test to Adaptive Testing

Lida Chen; Catherine Welch, University of Iowa; Stephen B. Dunbar, University of Iowa; Yibo Wang

This study compared three adaptive testing modes, including computerized adaptive testing (CAT), multistage testing (MST), and on-the-fly-MST by shaping (MST-S), in measurement accuracy and item exposure. Different designs were compared within each mode. Results suggested MST-S be used in practice due to its flexibility balancing measurement accuracy and item exposure.

Formative and summative uses for the log-linear cognitive diagnostic model (LCDM)

Zachary Conrad, USD 497; Peter Ramler

In today's standards-based educational environment it is imperative to conduct formative assessments throughout the year to track student progress as well as calibrate teacher-led, daily formative assessments with tests of known reliability. The log-linear cognitive diagnostic model (LCDM) provides output that fulfills these and additional diagnostic functions.

Imputations for Large-Scale Assessment Contextual Data: Is Recreation of Plausible Values Necessary?

Ting Zhang, AIR; Paul Bailey, AIR; Sinan Yavuz, University of Wisconsin-Madison; Huade Huo, AIR

The study tested two multiple imputation (MI) techniques for NAEP contextual data: (1) simple MI with existing plausible values and (2) the nested MI, in which plausible values are created conditioned on the imputed contextual dataset. The hypothesis is the nested MI produces more accurate estimates and variance estimations.

Investigating Production Rate of Short Essay Writing Using Large-scale Assessments

Tao Gong, Educational Testing Service; Mo Zhang, Educational Testing Service; Yang Jiang, Educational Testing Service; Chen Li, Educational Testing Service; Jiangang Hao, Educational Testing Service

Using two Winsight writing assessment datasets, we show that production rate, as a subskill of transcription, impedes writing process and performance: it correlates with individual properties of grade, gender, and ethnicity; and writing purposes mediate the thresholds of production rates as minimum requirement for good writing performance.

lz and lz* Person-Fit Statistics to Detect Aberrant Response Pattern

Yi Lu, Federation of State Boards of Physical Therapy; Yu Zhang, Federation of State Boards of Physical Therapy; Lorin Mueller, Federation of State Boards of Physical Therapy

This paper investigates two nonparametric person-fit statistics, lz and lz*. Research comparing the performance of these two statistics are sparse in literature. This study explores misfitting response pattern flagged by lz and lz* by specifying different estimation methods and different item response theory models in high-stake certificate testing context.

Measuring School-Level Traits using Multiple Individual-Level Informants

Tim Konold, University of Virginia; Elizabeth A. Sanders, University of Washington, Seattle

Methodological considerations in measuring and understanding the nature of informant differences (Level 1) when the target of measurement is at the organizational level (Level 2) are described in the context of a correlated trait – correlated method minus one (CT-C(M-1); Koch et al., 2015) latent model for multilevel applications.

Model Selection for Latent Dirichlet Allocation with Small Number of Topics

Constanza Mardones, University of Georgia; Jordan M. Wheeler, University of Georgia; Hye-Jeong Choi, University of Georgia; Allan Cohen, University of Georgia

Eight indices are studied to determine their accuracy for model selection for topic modeling with small numbers of topics. Preliminary results based on 30 replications suggest that Jensen-Shannon divergence and cosine similarity worked better than information criterion indices. A new interpretation is proposed for these latter methods.

Multistage Testing Design with Collateral Information Using Principal Components Analysis

Hyun Joo Jung

Han (2020) suggested a multistage testing design with intersectional routing approach (ISR) with regression models predicting initial scores. However, regression models change when different sections come first, which results in different final proficiency estimates. To address such issues, we propose ISR with principal components analysis and evaluate its performance.

Predicting Students' Future Performance Using Machine Learning Algorithms

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

In this study, we apply machine learning algorithms to explore the relationship between students' characteristics and performance using the PISA 2018 dataset. Multiple algorithms are applied for comparisons. Results show that XGBoost performs the best, with higher prediction accuracy. Socioeconomic status was the top feature contributing to the model.

Propensity Score Weighting with Principal Components: Inverse Probability of Treatment Weight

Yifang Zeng; Jaehoon Lee, Texas Tech University; Seungman Kim; Xinyang Li, Texas Tech University; Youngmin Kim, Texas Tech University

This simulation study examines the performance of using principal component (PC) scores for propensity score weighting (PSW), in comparison to the conventional approach of using all available variables to estimate PS. The results showed that PC-PSW has a great potential for reducing selection bias due to non-randomization in observational research.

The Multidimensional Item Response Tree Model and Its Application to Researching Response Styles

Biao Zeng, Beijing Normal University; Yanmei Li, Beijing Normal University; Hongbo Wen, Beijing Normal University

This study developed a multidimensional item response tree model and used this model to investigate the response style in the Undergraduate Learning Burnout scale. We found that participants showed a strongly avoid extreme response style and a significant interaction between learning burnout and response styles.

The validity of MCAT scores in predicting medical student performance on USMLE licensure exams

Kun Yuan, Association of American Medical Colleges; Cynthia Searcy; Andrea Carpentieri, AAMC

MCAT scores predict Step 1 and Step 2 Clinical Knowledge (CK) scores better than undergraduate grade point averages (UGPAs). On average, students with higher MCAT scores are more likely to pass licensure exams than those with lower MCAT scores and similar UGPAs. MCAT scores provide comparable prediction of licensure exam outcomes for students from different backgrounds.

Weights for the subdimensions in the Composite Model

Perman Gochyyev, University of California, Berkeley; Mark Wilson, University of California, Berkeley

Weighting schemes have been a central topic of discussions in studies focused on developing composite scores across multiple dimensions. We focus on the choices among various weighting schemes for the newly developed Composite Model. We review the literature then examine different weighting schemes using simulations and empirical data.

Focus on Linking and Equating
1:00 to 2:00 pm - Research Blitz Session

Chair:
Marianne Perie, Measurement in Practice, LLC

Participants:

Assessing the Impact of Equating Error on Mean Score Differences
Dongmei Li, ACT

Mean score differences are often used to monitor student progress over time. This study demonstrated that a simple statistic, that is, the standard error of equating sample mean score differences, can be used as a good estimate of the variability of group means due to random errors in test equating.

MIRT Observed Score Equating for CR Tests under the Nonequivalent Groups Design
Yoon Ah Song, Center for Applied Linguistics; Jiwon Choi, TEPS Center, Seoul National University

This study compares MIRT observed score equating to unidimensional IRT equating for CR tests under the common-item nonequivalent groups design when SR items were also used as external common items together. Results will be discussed in terms of equating accuracies from the real data and simulation study analyses.

An Approach to Test Equating Under the D-Scoring Method
Dimiter Milkov Dimitrov, National Center for Assessment; Dimitar V. Atanasov, New Bulgarian University

Under a recently developed method of test scoring and item analysis, called D-scoring method (DSM), item response functions are modeled on the D-scale (from 0 to 1) in classical and latent frameworks (Dimitrov, 2017, 2020; Dimitrov & Atanasov, 2020). This study offers an approach to test equating under the DSM.

Linking Two Vertically Scaled Tests of Interim and Summative Assessments
Jungnam Kim, NWEA; Hongwook Suh, Nebraska Department of Education; Nisha Padminiamma; Melinda Montgomery, NWEA; Christina Schneider, NWEA

The purpose of this study is to investigate the effect of these vertical scaling factors in linking two vertically scaled assessments of interim and end-of-year summative tests: whether one linking constant across grades is reasonable, if fixed method is practically sound, and how closely two tests should be administered.

When to Use Synthetic Linking Functions in Small-Sample Equating
Kylie N. Gorney, University of Wisconsin-Madison

In small-sample equating, a synthetic linking function may be used to reduce random equating error. However, this benefit comes at a cost: an increase in bias when test forms differ in difficulty. This study aims to identify which, if any, situations are best handled by synthetic equating.

A Comparison Study of Linking Methods Including Measurement Alignment for Mixed-Format Tests
Seongeun Kim, University of North Carolina; Kyung Yong Kim, University of North Carolina at Greensboro

This study introduces a measurement alignment as a viable alternative linking method and compare it to three other linking methods which was widely used. The feasibility of the measurement alignment method is assessed through simulation under various study conditions, including mixed format tests with both MC and CR items.

A Comparison of Anchor Lengths and Item Selection Methods in Small-Sample Equating

Kylie N. Gorney, University of Wisconsin-Madison

Two factors known to affect the quality of an equating anchor are length and item selection method. This study considers the way in which both factors affect equating results when samples are small, particularly when identity equating, nominal weights mean equating, synthetic equating, and circle-arc equating are used.

Focus on Adaptive Testing

1:00 to 2:00 pm - Research Blitz Session

Chair:

Tracy Gardner, Classic Learning Test

Participants:

Applicability and Efficiency of a Computerized Adaptive Test for Risk Assessment

Cihan Demir; Brian French, Washington State University

Youth risk measures can be cognitively and time burdensome given the number of items on the assessment. Assessments with fewer items may relieve this burden. A computer adaptive test (CAT) was simulated for a state-wide youth risk assessment. Results support a CAT version with little loss of measurement precision.

Effects of Splitting Testlets for Testlet-Based Computerized Adaptive Tests

Unhee Ju, Riverside Insights; Rong Jin, Riverside Insights; JongPil Kim, Riverside Insights

There are efficiency and adaptivity concerns with testlet-based CATs that have longer testlets (i.e., many items per testlet). This study examined the effects of splitting longer testlets into two shorter testlets on the performance of a simulated CAT moderated by testlet-selection methods and test lengths using empirical item pools.

Empirical Investigation of the Lexile Framework from an ELA CAT Assessment

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

Capitalizing on a unique CAT test design for an interim GK-HS ELA assessment that has embedded Lexile-type items, this study investigates a series of questions regarding how Lexile Measure scores compare against their empirically IRT calibrated counterparts. Results indicate there exists immense noise and potential bias in reported Lexile scores.

Effects on Item Parameter Estimates under a Multistage Testing

Akihito Kamata; Chalie Patarapichayatham, Southern Methodist University; Gonca Usta, Southern Methodist University

This study investigated how item parameter estimates were affected by a variant of multistage testing design, where students do not take more difficult items unless students provide correct responses. The results demonstrated that item discrimination parameters were substantially underestimated, while item difficulties were estimated without much bias.

Understanding Different Ways to Compute Measurement Errors and Score Reliability for Adaptive Tests

Yiqin Pan, University of Wisconsin-Madison; Lee Sung-Hyuck, Graduate Management Admission Council; Kyung (Chris) Han, Graduate Management Admission Council

This study explained the differences among various methods for computing standard error of measurement (SEM) and among score reliability indices. We compared their differences across conditions with varied test lengths and different true score settings. Findings from simulation and empirical studies offered guidelines for measuring and interpreting SEM and reliability.

An Invariance Preserving Cognitive Diagnostic Computer Adaptive Testing Algorithm

Yu Bao, James Madison University; Matthew James Madison, University of Georgia

Computerized adaptive testing algorithms rely on an assumption of item invariance. For diagnostic classification models, item invariance states that an examinee's attribute mastery is independent of the administered items. This study proposes a DCM-CAT algorithm to preserve this property. Through simulation, we evaluate the new algorithm under various test conditions.

Incorporating Response Choice in CD-CAT with a Multidimensional Bayesian Nominal Response Model

Catherine Elizabeth Mintz, University of Iowa; Jonathan Templin, University of Iowa

Partial knowledge is contained in item distractors, the choice of which can be modeled via nominal response models. However, these models are often overlooked in adaptive testing. Extending recent work, the current study examines the performance of a nominal response Diagnostic Classification Model in multidimensional adaptive assessment.

Item Evaluation Strategies

1:00 to 2:00 pm - Paper Session

Chair:

Xaviera Gonzalez-Wegener, UCL Institute of Education

Participants:

An Integrated Hierarchical Speed Accuracy and Answer Change Model

Aaron Myers, University of Arkansas; Irina Grabovsky, National Board of Medical Examiners

We are investigating the complexity of examinee response behavior in CBT using process data. The proposed model combines response accuracy, response times, item revisits, and answer changes. Person-level covariates are incorporated to investigate individual differences and interactions with items. An empirical example illustrates the model. Implications are discussed.

Instructional Sensitivity Indices for Ordinal Achievement Test Items

Anne Traynor, Purdue University; Xiaorui Li, Purdue University; Shuqi Zhou, Purdue University; Sandra Liliana Camargo Salamanca, Purdue University

For assessment scores to provide information about student learning progress, item responses must be affected by differences in instruction. We compare three instructional sensitivity indices for polytomous assessment items, including an item difficulty difference index derived from the generalized partial credit model, using simulation and analysis of science test data.

Predicting Problematic Items Using a Linguistic Complexity Framework: Findings From Cognitive Interviews

Kevin Close, Arizona State University; Yi Zheng, Arizona State University

In this study, we examine nuances of language-based construct-irrelevant variance by conducting in-depth cognitive interviews with 22 Spanish and Mandarin Chinese speaking students. A framework for linguistic complexity successfully predicted non-problematic items. Additionally, findings indicate that adding real-life context, and hence adding more words, may be unfair to such students.

Discussant:

Anthony Albano, University of California, Davis

(Invited Session) The value of assessment data from spring 2021: A debate

1:00 to 2:00 pm - Organized Discussion

February 23, 2021. Nearly a year ago, COVID-19 led to widespread school closures and a Department of Education blanket waiver of all statewide testing requirements. In fall 2020, the outgoing Trump administration resisted calls to cancel testing, while the incoming Biden administration remained silent on the issue. Since then, state officials, testing experts, teachers' unions, and civil rights advocates weighed in with calls to keep, limit, or waive state testing requirements. The U.S. Department of Education released a letter on February 22 outlining assessment expectations for states in 2021. The letter emphasized the importance of finding ways for states to test students even if that meant extending testing windows, relying on remote administration, or shortening the test. The Department declined blanket assessment waivers, but signaled openness to working with states needing additional flexibility. As we approach testing season, state leaders are wrestling with new administration logistics, analysis, and reporting challenges. How they address these challenges in the weeks ahead will add to the ongoing discussion about the value of testing in this unprecedented school year. This session will feature two teams of measurement leaders debating the value of state summative assessment during the 2020-2021 school year.

Session Organizer & Chair:

William A. Lorie, Center for Assessment

Presenters:

Scott Marion, Center for Assessment
Andrew Porter, University of Pennsylvania
Lorrie Ann Shepard, University of Colorado Boulder
Jon S. Twing, Pearson

Focus on Students with Disabilities

1:00 to 2:00 pm - Paper Session

Chair:

Leah Feuerstahler, Fordham University

Participants:

The I-SMART project: Empirical map validation

Jeffrey Hoover, University of Kansas; William Jacob Thompson, University of Kansas; Brooke Nash, University of Kansas; Jennifer Kobrin, ATLAS: University of Kansas

This study examines one piece of validity evidence for the learning map models used in the Innovations in Science Map, Assessment, and Report Technologies (I-SMART) project. Between-node correlations and estimated mastery profiles provided empirical support for the nodes and connections in the learning map models underlying the I-SMART assessment.

A Framework for the Evaluation of Assessment Accommodations

Maura O'Riordan, University of Massachusetts Amherst; Chris Domaleski, Center for Assessment

This research seeks to bridge the gap between the “best practices” for selecting, administering, and documenting accommodations with the practicality of providing such evidence through the development of a framework which can be used to determine and plan the evidence needed based on specific claims.

Identifying Constructs of the Transition Planning Process for Students with IEPs

David Johnson, University of Minnesota; Yi-Chen Wu, University of Minnesota; Ernest Davenport, University of Minnesota; Martha Thurlow, National Center on Educational Outcomes

This study examined the underlying factor structure of the IEP/Transition planning process for students with IEPs using the National Longitudinal Transition Study–2012 (NLTS 2012) dataset. Results identified four factors for students with disabilities—Youth/Parent Planning Invitation, Youth/parent Attendance, Youth Contribution, and Postschool Transition Planning.

Methods for Improving Validity for Cognitive Labs Using Purposeful Sampling Procedures

Melissa L. Gholson, Educational Testing Service; Jonathan Steinberg, ETS; Traci Albee, California Department of Education

Cognitive labs are a frequently used methodology in assessment development research and are particularly useful when examining low incidence populations where large pilot data collections are not feasible. This study used a new preliminary survey approach to purposeful sampling designed to inform stratification for a future K-12 state assessment.

On a Study of Group Invariance

Ernest Davenport, University of Minnesota; David Johnson, University of Minnesota; Yi-Chen Wu, University of Minnesota; Martha Thurlow, National Center on Educational Outcomes; Xueqin Qian, University of Kansas; Cynthia Matthias, University of Minnesota; John LaVelle, University of Minnesota

This study uses data from the National Longitudinal Transition Study 2012 to ascertain whether aspects of the transition meeting are consistent across IEP groups. Results suggest the experience was inconsistent and varied by data feature. These results have implications for group differences and how one aggregates data.

Discussant:

Danielle Guzman-Orth, Educational Testing Service

Automatic Item Generation Considerations
1:00 to 2:00 pm - Paper Session

Chair:
Guher Gorgun, University of Alberta

Participants:

Calibrating Automatically-Generated Items: A Comparison of Conventional and Hierarchical IRT Models
Mina Lee, University of Massachusetts Amherst; Scott Monroe, University of Massachusetts Amherst; Issac I. Bejar; Jonathan Weeks, Educational Testing Service; Ted Daisher, University of Massachusetts Amherst

In this study, constructed responses to 50 automatically-generated items were calibrated using different IRT models that make various assumptions about item isomorphicity, following Sinharay and Johnson (2008). The results from calibrations using each model and corresponding score estimates were compared. Results suggest conventional calibration of all items may be unnecessary.

Applying Weak Theory to Automatic Item Generation in CAT: A Case Study
Yanyan Fu, GMAC; Edison M. Choe, Graduate Management Admission Council; Jaehwa Choi, George Washington University; Hwanggyu Lim, Graduate Management Admission Council

This AIG case study applied weak theory to generate isomorphic items, or unique instances with equivalent psychometric properties. Three instances were generated from each of 25 models and pilot-tested in an operational CAT. DIF analysis will be used to check the equivalency of item parameters of instances within each model.

Strategies for practical implementation of Automatic Item Generation: Considerations of family level variance
Andrew Dallas, National Commission on Certification of Physician Assistants; Joshua Goodman, NCCPA

The current study investigates the relationship between family level variance and score precision. Using a simulation study, the researchers explore the conditions under which family level statistics can be used for form assembly and scoring without adversely impacting score precision.

Discussant:
Kirk Becker, Pearson

Modeling Response Time: A Collaborative Case Study on a High-Stakes Admission Exam

2:15 to 3:45 pm - Coordinated Paper Session

With the rapid development of technological infrastructure, computer-based testing is fast becoming the prevailing mode of test delivery. Consequently, item response times (RT) are now routinely recorded and analyzed for various purposes, including but not limited to checking speeded responses, detecting aberrant test-taking behaviors, and flagging potentially compromised items. To facilitate such analyses, numerous RT models have been proposed in literature and implemented by researchers over the years. However, virtually every model makes certain assumptions that may not always hold true in operational practice, thereby seriously challenging model fit to empirical data at large. Therefore, we propose a collaborative exercise in which four independent research groups each attempt to explain a particular set of RT data using their model of choice. The data come from a high-stakes graduate business school admission exam with a linear-on-the-fly testing (LOFT) design.

Session Organizer:

Edison M. Choe, Graduate Management Admission Council

Participants:

Joint Modeling of Responses and Response Times in LOFTs with Testlets

Hong Jiao, University of Maryland; Xin Qiao; Jung-Jung Lee

Mixture Response Time Model to Detect Aberrant Behaviors and Explain Item Nonresponses

Jing Lu, Northeast Normal University; Chun Wang, University of Washington

Utilizing Response Time to Measure Person Slipping in High-Stakes Tests

Yang Du; Justin L. Kern, University of Illinois at Urbana-Champaign

A Machine Learning Approach to Modeling Response Times

Yiqin Pan, University of Wisconsin-Madison

Discussant:

Hwanggyu Lim, Graduate Management Admission Council

Developing Successful and Impactful Assessment Products – Balancing Research and Business Considerations (Joint Session with Association of Test Publishers)

2:15 to 3:45 pm - Organized Discussion

For those working in the assessment industry, there are many competing demands that require constant attention. In many scenarios, these include demands on the time required to complete a project, the financial requirements of the project, and the need to develop or maintain assessments that are consistent with professional standards. During this session, a panel of seasoned measurement and educational technology professionals will discuss scenarios that require excellent judgment and experience to determine how to best meet these competing demands. Join us for this session and jump in and share your experiences attempting to juggle requirements to meet professional standards within the practical realities of the world.

Session Organizer:
Ada Woo, Ascend Learning

Chairs:
Jerry L. Gorham, Ascend Learning
Ada Woo, Ascend Learning

Presenters:
Wayne J. Camara, LSAC
Susan Davis-Becker, ACS Ventures, LLC
William Harris, Assoc. Of Test Publishers
John Weiner, PSI Services, LLC

(Invited Session) Assessment Literacy: Practical Applications and Implications (National Association of Assessment Directors)

2:15 to 3:45 pm - Coordinated Paper Session

During the 2019 NCME-NAAD symposium in Toronto, a panel of noted assessment literacy experts and district-level practitioners reached consensus: NAAD should launch an assessment literacy campaign focused narrowly on assessment directors. Raising levels of assessment acumen and skill, presumably, will increase the ability of assessment directors to manage and use their assessment systems effectively for a variety of purposes – e.g., to support teaching and learning, to monitor and evaluate program quality and equity, to influence assessment and accountability policy and practice, etc. Gradually, as the campaign gains traction and momentum, NAAD can broaden the effort to additional stakeholder groups - district leadership, central office staff, teachers and school administrators, students and parents, etc. Accordingly, the 2021 NCME-NAAD symposium constitutes a sort of interim report in a hybrid format that combines a coordinated paper session with an organized discussion. A "live" panel dialogue, moderated by Rick Stiggins, will follow a series of pre-recorded presentations. Some presenters will directly address particular aspects of assessment literacy, per se, while others will more obliquely address topics related tangentially to assessment literacy. We will encourage audience members to submit questions or comments via the Zoom chat feature.

Session Organizer:
Darin Kelberlau, Millard Public Schools

Chair:
Richard Stiggins, Assessment Training Institute

Participants:

Assessment Literacy: The Assessment Director's Role
Elda Garcia, National Association of Testing Professionals

Improving Student Writing: Assisted Writing Feedback Tools and Opportunities
Aigner Picou, The Learning Agency

Making Michigan the State of Assessment Literacy: Multiple Approaches to Promote Assessment Literacy
Edward Dean Roeber, Michigan Assessment Consortium

Improving Teacher Understanding and Use of Summative Assessment Data
Jeffrey Smith, Township High School District 214

Going beyond Scores: Understanding Response and Process in Large-scale Assessments

2:15 to 3:45 pm - Coordinated Paper Session

The transition from paper-based to digitally-based assessments in large-scale educational programs (e.g., National Assessment of Educational Progress (NAEP)) has provided unique opportunities to capture and analyze not only students' responses (e.g., finally submitted answers) but also their behaviors (e.g., a series of drag-and-drop actions) that lead to recorded responses. Methods to investigate such process data require necessary modifications to obtain useful inferences about students' (meta-)cognitive processes and problem-solving strategies beyond score-based analyses to better understand what students know and can do in those digitally-based assessments. Based on both the response and process data obtained from the science and math tasks from the NAEP assessments, this coordinated paper session highlights some recent studies that are designed to investigate students' response strategies, drag-and-drop action sequences, and knowledge levels leading to the correct or incorrect answers. Using both the "top-down" theoretical and "bottom-up" data-driven approaches, we will illustrate how to analyze and visualize process data and what response and control-of-variable strategies can be inferred by analyzing the drag-and-drop actions shown in the process data.

Session Organizers:

Burcu Arslan, Educational Testing Service
Yang Jiang, Educational Testing Service
Gary Feng, Educational Testing Service
Christopher Agard, Educational Testing Service

Chair:

Tao Gong, Educational Testing Service

Participants:

Visualizing Drag and Drop Action Sequences using Sankey Diagrams

Tao Gong, Educational Testing Service; Gary Feng, Educational Testing Service; Christopher Agard, Educational Testing Service; Gabrielle Cayton-Hodges, Educational Testing Service; Luis Saldivia, ETS

Understanding Fourth-Graders' Scientific Inquiry Practices with Process Data

Burcu Arslan, Educational Testing Service; Tao Gong, Educational Testing Service; Gary Feng, Educational Testing Service; Christopher Agard, Educational Testing Service; Madeleine Keehner, Educational Testing Service

Gaps between Knowing and Doing in Scientific Inquiry Practices within Large-Scale Educational Assessments

Yang Jiang, Educational Testing Service; Tao Gong, Educational Testing Service; Burcu Arslan, Educational Testing Service

Discussant:

Jesse Sparks, Educational Testing Service

Fostering Assessment Quality: Learning from Federal "Peer Review" Criteria, Process, and Impact

2:15 to 3:45 pm - Coordinated Paper Session

What characterizes quality in tests and testing, and how can that quality be fostered? The federally mandated "Peer Review" of state assessments is a quality control process with specific criteria that will be examined in this session. To provide all attendees essential background, this session will have an initial presentation on Peer Review, which states have undergone for nearly two decades. Then a panel of experts will discuss the nature, impact, strengths and weaknesses, and future of Peer Review, including ways in which evidence might be adjusted for school interruptions like those from COVID-19. The panel includes persons with expertise in assessment validity, assessment of special populations, a state assessment veteran, a consultant who has helped states comply with Peer Review and other technical criteria for over 20 years, and a U.S. Department of Education officer with responsibilities for Peer Review. Panelists will highlight the ways in which peer review bridges research and practice, the theme of the conference. The panel format will be adaptable to either on-site or virtual format, and will promote interaction and insight among the panelists, and between the panelists and the audience. Ample time will be allowed for audience participation in the discussion as well.

Session Organizer:

Liru Zhang, Assessment Consulting Services

Chair:

Brian Gong, Center for Assessment

Participants:

Peer Review: Past, Present, and Future

Donald Peasley, U.S. Department of Education

Peer Review: Influences on Professional Standards in Educational Measurement

Stephen Sireci, University of Massachusetts Amherst

Peer Review: Impacts on High-Stakes State Assessments

Liru Zhang, Assessment Consulting Services

Peer Review: Policy and Practices in Assessments for Special Populations

Martha Thurlow, National Center on Educational Outcomes

Peer Review: Effects on State Tests and Testing

Brian Gong, Center for Assessment

Topics in Standard Setting

2:15 to 3:45 pm - Paper Session

Chair:

Bryan R. Drost, Rocky River Schools

Participants:

Understanding Panelists: Providing Validity Evidence in Educational Standard Setting

Julie Pointner, DRC; Ricardo Mercado, Data Recognition Corporation; Jessalyn Smith, DRC

While standard setting methods have been widely studied, there is little research on these individual panelists participating in standard setting. This study surveys panelists and examines the degree to which the four factors (cognitive, social, political, and emotional), discussed in the literature, influence panelists' cut score recommendations.

Virtual Standard Setting: Applying the Many-facet Rasch Measurement (MFRM) Model

Charalambos Kollias, National Foundation for Educational Research

Virtual standard setting allows for reliable and valid cut scores to continue to be set when face-to-face workshops cannot happen. This paper will report on the Many-facet Rasch measurement (MFRM) model used to analyse judgments, compare media, and set cut scores.

An Examination of the Impact of Item Type in Standard Setting

Janet Mee, NBME; Peter Baldwin, National Board of Medical Examiners

In standard setting, judges may find it challenging to make internally consistent judgments across SAQ and MCQ items. This study examines judges' capacity to make these kinds of cross-item-type judgments. Preliminary findings suggest that judges may be more capable of making internally consistent judgments across question type than previously thought.

A Regression Discontinuity Approach to Find a Reading Fluency Proficiency Standard

Leslie Vanessa Rosales De Veliz, Juarez & Associates

A regression discontinuity method was explored as a data-driven method for setting the reading proficiency standard of Guatemala. The method used in this research assumes that the cut score must impact the performance in a more complex skill that a child will acquire later in his school life.

Applying a Mixture Rasch Model-based approach to Standard Setting

Michael R Peabody, National Association of Boards of Pharmacy; Timothy Muckle, Board of Pharmacy Specialties

The subjective aspect of standard-setting methods is often criticized, yet data-driven standard-setting methods are rarely applied. This paper examines the application of a mixture Rasch model approach across several testing programs of various sizes as well as a comparison with traditional standard-setting methods.

Discussant:

Marianne Perie, Measurement in Practice, LLC

Grading and Raters

2:15 to 3:45 pm - Paper Session

Chair:

Martha McCall, McKinsey & Company

Participants:

It's About Time: Multilevel Analysis of Grading Time on USMLE® Patient Notes

Beth Perkins, James Madison University; Jerusha J. Henderek, NBME; Thai Quang Ong, National Board of Medical Examiners

We used multilevel modeling to examine variability in grading time of patient notes from a high-stakes performance assessment to contribute to the limited research in this area. Differences in grading time were found for rater specialty and gender, as well as the task characteristics generating the constructed response.

Evaluating Targeted Double Scoring for Performance Assessments Using Simulated Data

Jing Miao, Educational Testing Service; Wei Wang, ETS; Sandip Sinharay, ETS; Yi Cao, Educational Testing Service; Chris Kelbaugh, ETS; Sandip Sinharay, ETS

In a targeted double scoring procedure for performance assessment, a subset of responses receives an independent second rating if their first rating falls into a pre-identified critical score range. This study simulates the second rating for examinees who had only one rating. Then the full data (including examinees with double scores and examinees with one score and one simulated score) are used to evaluate the critical score ranges and the accuracy of classification.

Evaluating Fairness in Automated Scoring

Nikole Gregg, Cambium Assessment, Inc.; Mackenzie Young, Cambium Assessment; Susan Lottridge, Cambium Assessment

Fairness practices in automated scoring do not address how bias may be introduced across parts of an engine. We expand fairness practices by investigating bias across multiple parts of a deep neural network engine, including: spell correction accuracy, word mapping to embeddings, and differences in embedding space location across subgroups.

Examining the Impacts of Ignoring Rater Effects in Mixed-Format Tests

Wenjing Guo, University of Alabama; Stefanie A. Wind, University of Alabama

We conducted simulation studies to explore the impacts of ignoring rater effects on student achievement estimates. The results suggest that under most conditions, a model with rater effects yields more accurate student achievement estimates. Only under certain conditions does a model without a rater effect parameter produce more accurate estimates.

A Model-Data-Fit-Informed Approach to Score Resolution in Rater-Mediated Assessments

Stefanie A. Wind, University of Alabama; Angela Adrienne Walker, Emory University

We explore the use of model-data fit analyses to inform score resolution procedures for mixed-format assessments as a theory-driven alternative to rater-agreement-based approaches. We compared our approach to an agreement approach using a simulation study. The fit-based approach resulted in more-reasonable estimates for larger proportions of students.

Discussant:

Mark David Shermis, American University of Bahrain

Standard Setting Challenges and Solutions for Innovative Assessment System Designs

4:00 to 5:30 pm - Coordinated Paper Session

Innovative assessments designed to target deeper learning have gained traction in K-12 assessments. Standard setting—that is, setting cut scores to define performance levels—can be a challenge for innovative assessments. Current standard setting methodologies, many of which are best suited for traditional summative assessments with selected-response items, require selecting or adapting methods that are appropriate for innovative assessments. This session will provide practical guidance on adapting and applying current standard setting methods to meet the needs of three innovative assessment designs. The three assessments and standard-setting applications covered in this session span a wide range, including a phenomena-based writing assessment in science using the Body of Work method, a statewide science assessment containing item clusters using a modified Item Descriptor (ID) Matching method, and a performance assessment system developed under the IADA initiative using Contrasting Groups methodology. This symposium will inform practitioners on adapting standard setting methodologies while maintaining the validity of the methods and results.

Session Organizer:

Qi Qin, Gwinnett County Public Schools

Chair:

Steve Ferrara, Cognia

Participants:

Setting Performance Standards for Phenomena-based Writing Assessments using Body of Work Methodology

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

Modified ID Matching Standard Setting for Item Cluster Test Designs

Eric Moyer, Pearson; Jennifer Lynn Galindo, Pearson; Scott N. Strickman; Liru Zhang, Assessment Consulting Services

Applying Contrasting Groups Standard Setting Methodology to a Performance Assessment Program

Carla M. Evans

Discussant:

Susan Davis-Becker, ACS Ventures, LLC

Scrutinizing item responses and response times: Experimental and analytic approaches

4:00 to 5:30 pm - Coordinated Paper Session

In this session, we aim to make our measurements matter by studying experimental and analytic approaches for improving the use of both item responses and response times in educational assessments. In the first two papers, the focus is on results from an experimental study in which technology was leveraged to investigate the impact of different scoring rules (accuracy vs. speed and accuracy), timing conditions (no time limit vs. test and item limits), and feedback conditions (no feedback vs. accuracy and speed feedback) on item responses and response times. Response times are often used rather casually as collateral information, but the results from this experimental study show that it matters under which digital-based administration conditions they were collected and that this has an impact on joint modeling. In the second set of papers, analytic approaches are discussed to study the intricacies of conditional dependencies within and between item responses and response times in the context of digital-based large-scale educational assessments (e.g., NAEP and PISA). The extent to which conditional dependencies occur is studied as well as how to extend the standard latent regression item-response theory models to account for them.

Session Organizer & Chair:

Peter van Rijn, ETS Global

Participants:

Measurement invariance across different scoring and timing conditions in joint modeling of item responses and response times

Usama Ali, Educational Testing Service; Peter van Rijn, ETS Global

Effect of immediate feedback on performance in practice tests

Yigal Attali, Duolingo; Usama Ali, Educational Testing Service

Impacts of item types in the response-time conditional dependencies

Hyo Jeong Shin, Educational Testing Service; Paul Adrian Jewsbury, Educational Testing Service

Time-accuracy conditional dependencies and latent regression models

Paul Adrian Jewsbury, Educational Testing Service; Hyo Jeong Shin, Educational Testing Service; Peter van Rijn, ETS Global

Discussant:

Dylan Molenaar, University of Amsterdam

Suggestions for Fairness and Equity, as well as Quality, in Testing

4:00 to 5:30 pm - Coordinated Paper Session

In this session, we first provide a history lesson involving examples of tests contributing to negative impact. These examples will make the issues salient and help us learn from them. Next, a comprehensive framework anchored in evidence to examine fairness of assessments will be presented. Next, two presentations will be provided suggesting specific solutions to increase equity and quality of assessments. The first in this set will discuss enhancements to assessment design to improve equity in the content and constructs of tests. The second will discuss Positive Assessment principles for the use by test developers to ensure the quality and enhance the equity of the assessments. These solutions will be resented in a way to stimulate discussion and enhance assessment practices for the benefit of all involved. Finally, equity in testing will be defined in the context of the personal, cultural, and situational components of examinees that include aspects of opportunities to learn. Suggestions for practice will be provided.

Session Organizer:

Thanos Patelis, Fordham University, Teachers College, University of Kansas

Participants:

Some Historical Vestiges of Bias in Test Development

Kurt Geisinger, University of Nebraska-Lincoln

Foundational Concepts in Fairness in Assessment

Maria Elena Oliveri, Buros Center for Testing-UNL

Reimagining Construct Representation to Promote Equity in Principled Assessment Design

Kristen Huff, Curriculum Associates

Think Positive: How to Use Assessments to Bolster Student Learning

Stephen Sireci, University of Massachusetts Amherst; Sergio Araneda, University of Massachusetts Amherst

Equity in Assessment Goes Beyond the Instrument – the Context Matters

Thanos Patelis, Fordham University, Teachers College, University of Kansas

Electronic Board Session #4

4:00 to 5:30 pm - Electronic Board Session

Participants:

Achievement Gaps when NAEP-like Conditioning Method is Applied to ECLS-K 2011 Assessments

Soo Lee, American Institutes for Research; Burhan Ogut, American Institutes for Research; Markus Broer, American Institutes for Research; William C. Tirre, U.S. Department of Education

The conditioning model refers to a process that use both cognitive item responses and student's additional background information (e.g., student socioeconomic status; SES) in estimating scale scores. This study investigates whether achievement gaps in ECLS-K: 2011 assessments change if a NAEP-like "conditioning model" is applied in ability or score estimation.

A Comparison of Classification Methods in a Computerized Adaptive Test

Ozge Ersan, University of Minnesota Twin Cities; Joseph DeWeese, University of Minnesota Twin Cities

We compared the performances of three classification methods in a CAT. These methods are Confidence Interval, Sequential Probability Ratio Test, and Generalized Likelihood Ratio methods. Results suggest the advantages of GLR in terms of accuracy and efficiency.

A Confirmatory Restricted 4PNO Model

Justin L. Kern, University of Illinois at Urbana-Champaign; Steven Culpepper, University of Illinois at Urbana-Champaign

There is renewed interest in the four-parameter IRT model (4PM). However, the identifiability of the 4PM is questionable. One recent paper showed how to identify the 4PM, presenting an exploratory approach to meeting these conditions (Kern & Culpepper, 2020). In this project, we explore a confirmatory version of this model.

Approaches to Reduce Ability Differences of Equating Samples

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

This study compares five approaches for reducing group nonequivalence in an equating design when randomization is unsuccessful and there are few common items. Group adjustment through demographic data, a weak anchor, or a mix of both, is evaluated in terms of equating accuracy.

A Social Network Analysis of Answer Change Behavior Using NAEP Process Data

Xin Qiao; Juanita Hicks, AIR

Given that process data include rich information on students' assessment behavior it may shed new insight on the investigation of answer change. A social network analysis was used to highlight assessment behaviors collected using process data to further explore answer change patterns and behaviors.

Assessing Severity Effects on an Object Standard Setting Exercise: A Simulation Study

Karen Fong

The study investigated rater severity effects on the cut score location on an Objective Standard Setting exercise. Panelists rated items within their expertise, as opposed to the traditional practice of rating all items. Results show that severity effect slightly affected the pass rates associated with the upper bound cut scores.

Designing Mastery Assessment Item Pool for Online Hybrid Learning System

Jinah Choi; Audra Kosh, Edmentum, Inc.

When skill-mastery tests co-exist with summative computerized adaptive tests (CAT) within an online learning program, it is

important for scores from the mastery tests to be as valid/reliable as those from the CAT. This research examines designing skill-mastery item pools psychometrically in order to most efficiently/accurately assign students individualized online-curriculum.

Does decoding really affect reading? ——A test of Chinese decoding threshold hypothesis

JingYi Li, Beijing Normal University

Based on the Lexical Quality Hypothesis and the particularity of ideograph, this study proposes a Chinese decoding threshold interval hypothesis, reconciling two contradictory models in the Simple View of Reading. This hypothesis is identified by cross-sectional data and examined by longitudinal data, which also has implications for reading practice.

Evaluating Field Testing in Multistage Testing in a Large-Scale Language Assessment

Kyoungwon Bishop, WIDA at UW-Madison; Sakine Gocer Sahin, WIDA at UW-Madison

This study aims to evaluate field testing practice in a multistage test of a large-scale English language assessment and to find optimal calibration design. To maintain measurement scale between operational and field test items properly, this study addresses location and distribution of field test items, and sample proportions for calibration.

Evaluation of Routing Decision in Testing: Logistic Regression and Signal Detection Theory

Sakine Gocer Sahin, WIDA at UW-Madison; Kyoungwon Bishop, WIDA at UW-Madison

The purpose of study is to evaluate binary placement rules using logistic regression in light of signal detection theory in a large-scale English language proficiency test. Data-driven placement decision was examined to explore the extent and the direction of bias.

Exploring the relationship between item response time and item characteristics

Aijun Wang, FSBPT; Yu Zhang, Federation of State Boards of Physical Therapy; Lorin Mueller, Federation of State Boards of Physical Therapy

Selection of items with known response time (RT) help test developers set proper time limit and eliminate the extraneous variances in test scores. This study tries to explore how item characteristics are related to RT before the items are field tested and utilize the relationship in test development.

Factor Analysis of Ordinal Data and the Number of Response Categories

Mohammed Abulela; Amaniel Mrutu; Ernest Davenport, University of Minnesota

We investigate the effect of the number of response categories on several exploratory item factor analysis procedures. Data were simulated for various conditions including: simplicity of factor structure, factor correlations, sample size, and number of response categories. We used several criteria to investigate fit including RMSE, AIC, bias, etc.

Inculcating intellectual character: Pilot results of an online module to enhance undergraduate intellectual virtue at university

Gabe Avakian Orona, University of California, Irvine; Duncan Pritchard, University of California, Irvine

The concept of intellectual virtue, and its renewed relevance for undergraduate education, has surfaced as a noteworthy framework in situating the development of 21st century competencies. Based on these data, students are satisfied with the module and show growth in intellectual curiosity.

Inter-Rater Reliability of Evaluators Judging Teacher Performance: Alternatives to Cohen's Kappa

Richard Lambert, UNC Charlotte; Timothy Scott Holcomb; Bryndle L Bottoms, University of North Carolina at Charlotte; Kawanna Jackson, UNC Charlotte

Questions persist about the validity of Kappa coefficients when prevalence of specific rating scale categories is low and agreement is high. Teacher evaluation data confirmed both the shortcomings of Kappa and the robustness of Gwet's AC1 and the Lambda Coefficient of Rater-Mediated Agreement relative to these problematic data conditions.

Investigating DIF of Items in a CAT Test between States and Grades

Siyu Wan, University of Massachusetts Amherst; Yeow Meng Thum, NWEA

This study investigated the score invariance of a widely used CAT assessment related to states and grades. DIF analyses for items representative of the bank were performed by using the logistic regression procedure on raw and summarized dataset. The finding provided evidence that items operated similarly across grades and states.

Investigating the multidimensionality of cultural resiliency in a multicultural environment

Alejandra Miranda; Mireya Carmen-Martinez Smith, University of Minnesota

Nowadays, inequalities have been worsened by COVID-19 and racial injustice. Promoting 21st century skills has become crucial for youth. Using student data, we assess the dimensionality of a cultural resilience measure. This validity study provides evidence about how to use this tool, bridging the gap between research and practice.

Investigating the Relationship between Test Information and Routing Accuracy in an MST

Shumin Jing; Louis Roussos, Cognia; Liuhan Cai, Cognia

This research examines the relationship between Stage 1 test information and the probability of correct routing from Stage 1 to Stage 2 under a 1-3 multistage test design. Based on the results of real data analyses, a simulation study is conducted and a comprehensive evaluation of the results is presented.

Literature Review of Culturally Responsive Assessment and Educational Practices

Sandra Margaret Botha

This review summarizes literature relevant to culturally responsive theory and practices in assessment and education. Topic-related research studies and papers were selected through database searches and snowballing.

Paternal Incarceration: Patterns and Achievement in PIAAC US Population and Prison Study

Carina M. McCormick, Buros Center for Testing

The disproportionate incarceration of parents of minority students is suspected to relate to gaps in child educational outcomes, but current research is sparse. This study compares key educational data for fathers and non-fathers in prisons and in the U.S. population, by racial/ethnic group, using PIAAC and the PIAAC Prison Study.

Tiered Claims: A New Approach to Claims about Students in NGSS Assessment

Sanford Student, University of Colorado Boulder; Brian Gong, Center for Assessment

We introduce tiered claims, a novel approach to structuring student-level claims in large-scale assessment relative to complex or multidimensional domains, with the NGSS as a guiding example. We discuss implications for item and test development; the connection of psychometric analysis to validity arguments; and alignment, standard setting and reporting.

True Q-matrix estimation for conjunctive attribute space using regularization techniques

Jihang Chen, Boston College; Zhushan Mandy Li, Boston College

As the increasing needs of cognitive assessment in psychological and educational measurement for assessing students' mastery of skills, Q-matrix identification became critical to specify the item-attribute relationship. We aimed to explore the stability and accuracy of Q-matrix estimation using different models under different conditions through the regularized maximum likelihood.

Use of Response Times and Person-Fit Statistics to Detect Possible Item Pre-knowledge

Nooree Huh, ACT; Chi-Yu Huang, ACT; Yang Lu, ACT

This study evaluated the potential usefulness of the combination of item response times and person-fit methods in detecting possible cheaters in online testing. The examinees' ability levels, the number of breached items, and the test lengths, were examined in the study.

Validity Evidence for State Summative Assessment Programs

Teresa Dawber, Council for Aid to Education; Joanna Tomkowicz, Data Recognition Corporation

The study meta-analyzes sources of validity evidence presented in technical reports for state summative assessment programs. Early results indicate variability of breadth and depth of evidence based on Test Content, Internal Structure, Response Processes, Relations to Other Variables, and Consequences of Testing for the surveyed testing programs.

What is Mastery in MIRTMs: Connection between CDMs and MIRTMs

Mingqi Hu, University of Illinois at Urbana-Champaign; Jinming Zhang, University of Illinois at Urbana-Champaign

Multidimensional IRT models (MIRTMs) measure multidimensional latent abilities, which shares similarities with CDMs. The study connects them to see if MIRTMs can provide some diagnostic information. Cutoff point is used for transformation. In simulation, MIRT estimates in CDAs may perform similar results as CDM after transformation and provide precise information.

Diagnostic Assessments: Moving from Theory to Practice

4:00 to 5:30 pm - Coordinated Paper Session

In recent years, there has been a call for assessments to provide increasingly detailed and actionable scores, while simultaneously decreasing overall testing time. This demand is an incredible challenge for the educational assessment community, but one that is answerable through the use of diagnostic assessments and diagnostic classification models (DCMs). Despite these benefits, DCMs have not been widely adopted for use in operational settings. This session ties together four papers that describe, in practical terms, how to design, implement, and support the use of DCM-based diagnostic assessments for operational use. The first presentation illustrates how assessments and items can be designed to elicit fine-grained diagnostic information about students, rather than assessing a single latent trait. The second presentation discusses the decision-making process involved with DCM model building, model selection, and practical model fit considerations. The third presentation illustrates how the scores from a diagnostic assessment can be reported in a meaningful way to support actionable next steps. The fourth presentation describes how traditional psychometric methods can be revised in order to provide technical documentation that is required of any operational assessment. The session ends with commentary from a national expert in diagnostic models and their use in applied settings.

Session Organizer & Chair:

W. Jake Thompson, University of Kansas

Participants:

Designing a diagnostic assessment

Leanne Ketterlin Geller, Southern Methodist University

Weighing parsimony and flexibility in diagnostic classification model selection

Meghan Fager, National University; Matthew James Madison, University of Georgia

Communicating results of diagnostic assessments

Laine Bradshaw, University of Georgia

Technical evidence for diagnostic assessments

W. Jake Thompson, University of Kansas; Amy Clark, ATLAS: University of Kansas; Brooke Nash, University of Kansas

Discussant:

Robert Henson, University of North Carolina

Topics in Measuring Growth

4:00 to 5:30 pm - Paper Session

Chair:

Luciana Cancado, Curriculum Associates

Participants:

Growth Measure Accuracy in a Learning Progression Framework: A Simulation Study

Duy N. Pham, Educational Testing Service

This study examined the accuracy of a growth measure based on learning levels of a learning progression. Learning levels were generated using the Rasch model; the model was then fit to the simulated data. The results indicated that 53 to 78 percent of true growth could be recovered.

Growth Measure Comparisons of Vertical Scales Versus Grade-Level Scales

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

The study proposes Grade-Level Scaling, a new scaling methodology for measuring student's growth, by adapting the advantages of vertical and horizontal scaling. Grade K-8 items are calibrated on one vertical scale and 9 Grade-Level scales with empirical CAT assessment data. Student growth measure is compared using the two scaling methods.

Measuring Growth Using Accelerated Longitudinal Designs for Linking Multiple Age-Cohort Growth Curves

Yeow Meng Thum, NWEA

Accelerated longitudinal designs (ALDs) are explored for estimating mathematics growth over seven grade levels, when only three years of student longitudinal data nested within schools are available for each cohort. Implications from sensitivity to multilevel growth model specification for mean and covariance structures are examined for establishing growth measures.

Student Growth Percentile and Latent Change Scores with Two Time-Points

Dakota Wayne Cintron, University of Connecticut; Nina Deng, Kaplan INC.

Measuring change with two time-points is disputable. With two time-points, student growth percentile measures change relative to peers whereas latent change scores are a measure of latent change. This study investigates their performance in a simulation using magnitude of change, test length, and sample size as design factors.

Models for Aggregate Growth and Progress Using Multiyear, Multicohort Datasets

Benjamin R. Shear, University of Colorado Boulder; Andrew Ho, Harvard Graduate School of Education; Sean Reardon, Stanford University

We propose a 4-level longitudinal growth model to summarize multiyear, multicohort student test score data for educational achievement monitoring. We illustrate use and interpretation of the model, evaluate reliability of the parameter estimates, and discuss extensions of the model including covariates and estimation when only aggregate data are available.

Discussant:

Chris Domaleski, Center for Assessment

Techniques in Machine Learning or Artificial Intelligence

4:00 to 5:30 pm - Paper Session

Chair:

Leslie Keng, Center for Assessment

Participants:

Building Knowledge Components Network from Student Performance: A Collaborative Filtering Approach

Shuai Zhu, TAL Education Group; Kaifu Wang, TAL Education Group; Shouye Peng, TAL Education Group; Yuying Ji, TAL Education Group

Experienced educators might be able to subjectively determine the relationships between different knowledge components, but it's hard for them to finely quantify those relations. Inspired by the recommender system, we propose a collaborative filtering approach to discover the quantitative relations among knowledge components to construct a knowledge components network (KCN).

Propensity Score Estimation: Comparison of Logistic Regression, DNN, and CNN

Seungman Kim; Jaehoon Lee, Texas Tech University

This study proposes two machine learning techniques—deep neural network (DNN) and convolutional neural network (CNN)—as new estimators of propensity score that can algorithmically handle nonlinear relationships and interactions of covariates. Simulation was conducted to examine the performance of DNN and CNN in comparison to the conventional method.

LSTM-MTSL: A Deep-Learning Approach to Multivariate Time-Series Log-event Prediction of Academic Performance

Chang Lu, University of Alberta; Maria Cutumisu, University of Alberta

This study proposes LSTM-MTSL, a deep-learning Long Short-Term Memory approach to Multivariate Time-Series Log-event prediction of 367 undergraduate students' academic achievement based on Moodle log files. Results show that LSTM-MTSL outperformed multiple linear regressions and artificial neural networks and made early accurate predictions based on the first week.

Semi-Automatic Scoring of Constructed Responses in Large-Scale Assessments

Nico Andersen, DIPF | Leibniz Institute for Research and Information in Education; Fabian Zehner, DIPF | Leibniz Institute for Research and Information in Education, Centre f. Int. Student Assessm.; Frank Goldhammer, DIPF | Leibniz Institute for Research and Information in Education, Centre f. Int. Student Assessment

To minimize the human scoring effort in large-scale-assessments, we developed a computer-linguistic approach that dynamically supports human raters. This semi-automatic method was evaluated in simulations using responses from the PISA reading assessment and saved, on average, about 42% of the scoring effort with an accuracy of 98%.

Comparison of Human Rater and Automated Scoring of Test Takers' Speaking Ability and Classification Using Item Response Theory

Zhen Wang, Cambium Assessment

Automated scoring has been developed and has the potential to provide solutions to some of the obvious shortcomings in human scoring. In this study, we investigated whether automated scoring and a series of combined automated scoring and human scores were comparable to human scores for an English language assessment speaking test. We found that there were some systematic patterns in the five tested scenarios based on item response theory

Discussant:

Susan Lottridge, Cambium Assessment

Fireside chat with the Classroom Assessment Task Force
5:45 – 7:00pm – Discussion

The Task Force will be sharing news about its new status, and introducing new CATF members to the wider NCME community. The fireside chat will also be an opportunity for CATF to share up-dates about is various initiatives during the past year, including additions to the FACT website and details about the upcoming 4th biennial NCME Special Classroom Assessment Conference (virtual) in October 2021. Plans for journal special issues of recent Classroom Assessment Conference papers will be shared and brainstorming for the 2022 NCME Annual Conference pre-session with educators will be invited. We will encourage input from the NCME membership on these and other ideas for future initiatives.

Organizers & Hosts:
Caroline Wylie, Educational Testing Service
Alison Bailey
Members of CATF

NCME Yoga
7:30 to 8:30am

Rosemary Reshetar, National Conference of Bar Examiners

Come join your colleagues and friends in this all-level friendly hatha yoga class. The focus will be on practicing basic to intermediate poses, alignment principles, and breathing techniques. Modifications for multiple abilities will be offered throughout. The class will end with relaxation and you'll leave alert and refreshed for a full NCME conference day. I'll keep my camera on. You can choose to be on or off camera for the session.

Focus on English Language Learners

9:00 to 10:30 am - Paper Session

Chair:

Yeow Meng Thum, NWEA

Participants:

Validity Evidence for English Learners Testing with Supports on the ACT

Joann Moore, ACT; Dongmei Li, ACT; Yang Lu, ACT

This study evaluated the impact of providing testing supports for English learners taking the ACT® test. Modest score gains and stronger relationships with high school grades were found. Psychometric analyses did not find evidence of bias, suggesting that the supports are benefitting English learners without conferring an unfair advantage.

Does Time in ELL Programs Affect Properties of ELPA Vertical Scales?

Jon Lehrfeld, Educational Testing Service; Terran Brown, ETS

In constructing vertical scales for English language proficiency assessments, the assumption that older students have greater proficiency than younger students may not be accurate. We examine how the properties of one such vertical scale change when time spent in ELL programs is used as inclusion criteria to construct the scale.

Examining the relationship between technologies and second language learning outcomes: A meta-analysis

Songtao Wang, University of Victoria

This meta-analysis study examined the relationship between using technologies and second language learning outcomes. 30 experimental studies were included which yielded 43 effect sizes. Results showed a large effect size in favour of technology-integrated instructions with substantive between-groups heterogeneity. Practical implications on meta-analysis and L2 measurements were discussed.

Human Scoring versus Automated Scoring for EL Students in Statewide Writing Assessment

Yen Vo, University of Iowa; Heather Rickels, University of Iowa; Catherine Welch, University of Iowa; Stephen B. Dunbar, University of Iowa

This study examined EL and non-EL performance in a statewide writing assessment. After controlling for student characteristics with propensity score matching (PSM), DIF results indicated that paper format (scored by humans) tended to favor EL students. This trend was not found with online format (scored by automated scoring).

Examining Lexical Features of Standardized Math Test Items: A Text Mining Approach

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

English Language Learners exist in academic disparity to their monolingual peers, which is especially apparent through mathematical standardized testing. The language used in these word problems is often inappropriate and verbose. This study aims to create a corpus of standardized mathematical items and document the most used terms.

Discussant:

Robert Dolan, Diverse Learners Consulting

The Past, Present, and Future of Item Difficulty Modeling

9:00 to 10:30 am - Coordinated Paper Session

This 90-minute coordinated session will consist of four presentations focused on item difficulty modeling (IDM), beginning with a literature review that synthesizes findings from over 100 studies. The remaining presentations are instances of IDM studies that describe different modeling strategies applied to K-12 assessments and a college entrance exam. The first IDM study uses the linear logistic trait model with person and item covariates to predict the difficulty of elementary math items. This presentation also outlines a process that integrates IDMs into item templates for continual improvement. The second study compares alternative models, from the more conventional (multiple regression) to more contemporary modeling approach (machine learning algorithms), with special emphasis placed on the role of range achievement level descriptors as construct validity evidence. The final study compares different machine learning approaches to predict item difficulty for reading comprehension items in a college entrance exam. Methods included supervised machine learning using human-engineered item features, and a feedforward neural network coupled with a text encoder to automatically extract item features. To conclude the session, a discussant will provide thoughts on the various methods presented in the papers.

Session Organizer:

Garron Gianopulos, NWEA

Participants:

Item Difficulty Modeling: The State of the Art

Steve Ferrara, Cognia; Jeffrey Steedle, ACT; Roger Frantz, Questar

Integrating Item Difficulty Modeling into Test Design for Continual Improvement

Garron Gianopulos, NWEA; Jungnam Kim

Predicting Item Difficulty Using Item RALD Levels and Other Item Features

Jing Chen, NWEA; Christina Schneider, NWEA; Paul Nichols, NWEA

Predicting the difficulty of reading comprehension items from a college entrance exam

Brad Bolender, ACT; Shi Pu, Independent; Rick Meisner, ACT; Colin Dingler, ACT

Discussant:

Kristin M. Morrison, Curriculum Associates

Leveraging Response Process Data to Support Testing Programs: Strategies and Real-world Examples

9:00 to 10:30 am - Coordinated Paper Session

Digitally based assessments allow the capture of learners' response processes information at finer time granularity. Leveraging this additional information properly will help to improve assessments in various psychometric areas such as validity, reliability, comparability, and fairness (Ercikan & Pellegrino, 2017; Mislevy et al., 2014). Though a high-level value proposition of response process data is relatively straightforward to make, real complications and challenges come from the details of developing feasible pathways towards the materialization of the promises. In this coordinated session, we include five presentations, each of which shows a strand of research on how to use the response process data from a large-scale assessment at ETS to support the testing program in various ways. We hope these presentations can update the community of the current progress and practice around using response process data analytics to support large-scale testing programs. We are looking forward to feedback, discussions, or debates from the community to help us develop future research agenda.

Session Organizer:

Jiangang Hao, Educational Testing Service

Participants:

Data and methodological strategies for analyzing response process data in practice

Jiangang Hao, Educational Testing Service; Robert Mislevy, Educational Testing Service; Chen Li, Educational Testing Service

Exploring the progression of writing fluency in large-scale assessments using keystroke logs

Yang Jiang, Educational Testing Service; Jiangang Hao, Educational Testing Service

Assessment Platform Tool Usage Analytic

Jie Gao, Educational Testing Service

Exploring response time and re-visit pattern in student response process

Qiwei He, Educational Testing Service; Bingchen Liu

Using Artificial Intelligence for Constructed-Response Scoring: Some Practical Considerations

9:00 to 10:30 am - Coordinated Paper Session

This coordinated session includes 4 papers that explore applied aspects of using artificial intelligence (AI) for constructed-response (CR) scoring. Paper 1 (Raczynski, Choi, & Cohen) focuses on considerations for developing CR items that will be AI-scored. The authors use latent class analysis to identify and describe characteristics of CR items shown to be AI score-able. Paper 2 (Lottridge, Ormerod, & Jafari) and paper 3 (Wheeler & Cohen) explore alternatives to Latent Semantic Analysis (LSA) as a method of text analysis and AI score prediction. Lottridge et al. describe deep learning or multi-layer concurrent and neural networks methods for building an AI engine and an applied study that addresses some of the challenges associated with these methods. Wheeler and Cohen focus on Latent Dirichlet Allocation (LDA) and offer an empirical examination of how it compares to LSA as a method for analyzing written text. Paper 4 (Cohen & Levi) is an applied study of different rating protocols for operational scoring—those involving just human raters and those involving both human raters and AI—in the interest of reducing error variance. Mark Shermis of the American University of Bahrain will offer perspective as the Discussant.

Session Organizer:

Kevin Raczynski, University of Georgia

Participants:

Using Latent Class Analysis to Explore the AI Score-ability of Constructed-Response Items

Kevin Raczynski, University of Georgia; Hye-Jeong Choi, University of Georgia; Allan Cohen, University of Georgia

Explaining Scores Produced from Neural Net-Based Engines

Susan Lottridge, Cambium Assessment; Chris Ormerod, Cambium Assessment; Amir Jafari, Cambium Assessment

A Comparison of Latent Semantic Analysis and Latent Dirichlet Allocation

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

AES as an Aid in Quality Assurance of Essay Scoring

Yoav Cohen, National Institute for Testing; Effi Levi, National Institute for Testing & Evaluation

Discussant:

Mark David Shermis, American University of Bahrain

Topics in Item Response Theory

9:00 to 10:30 am - Paper Session

Chair:

Brian C Leventhal, James Madison University

Participants:

Precision-Weighted IRT Scale Transformations Via Response Function Methods

Alexander Weissman, Law School Admission Council; Wim J. van der Linden

This study extends the work of Barrett and van der Linden (2019) by incorporating IRT parameter estimation error into response function scaling, yielding precision-weighted linear transformation constants. This method is compared with Haebara's (1980) and Stocking and Lord's (1983) methodologies, and applied to automated assembly of anchor item sets.

2PL Model: Compare Generalized Linear Mixed Model with Latent Variable Model based IRT framework

Jihong Zhang, University of Iowa; Terry Ackerman, University of Iowa; Yurou Wang, University of Alabama

Recently fitting IRT models using the generalized mixed logistic regression modeling (GLMM), also called generalized latent variable modeling (GLVM), has become popular in large-scale assessment research because GLMM combines multilevel structural models with IRT measurement models. However, the estimation accuracy of item parameters between these two modelings is not well examined. This study aimed to compare the estimation results of the GLMM based 2PL model with the traditional IRT model under different sample sizes and test length conditions. The results showed that for both GLMMand IRT modeling, item threshold's estimates were more accurately estimated than item discrimination parameters. We also found that GLMM estimation had lower accuracy than traditional IRT modeling when estimating both high and low discriminated items.

Bayesian Estimation of Ability in Hybrid Item Response Models on Mixed-Format items

Jiawei Xiong, University of Georgia; Allan Cohen, University of Georgia; Xinhui Xiong, Educational Testing Service

In this paper, a Bayesian incorporated hybrid item response model is described for mixed-format items. Students' posterior ability is estimated using the prior ability distribution estimated from the multiple-choice items in the test. This model is compared with the graded response model through both the simulation and real data.

Outlier Detection Using Sampling Variance of IRT Parameter Estimates under 3PL Model

Chunyan Liu, National Board of Medical Examiners; Daniel Jurich, National Board of Medical Examiners

In equating practice, the existence of outliers in the anchor items may threaten the validity of test score interpretations. The current simulation demonstrates that the sampling variance of the IRT parameter estimates can be helpful in detecting true outliers and lead to improved equating accuracy and examinee ability estimation.

Modeling the Model Error as a Random Effect in IRT Models

Shuangshuang Xu; Yang Liu, University of Maryland, College Park

The proposed study provides a stochastic framework to specify model errors as random effects in IRT models, and investigates the model parameter recovery in the framework under different conditions. A bias correction is also used to adjust the discrepancy between the true and estimated parameters in the current approach.

Discussant:

Jessalyn Smith, DRC

The Future of K-12 Assessment: Is there One?

9:00 to 10:30 am - Organized Discussion

Prior to COVID-19, educational assessment was being challenged by its stakeholders on many fronts and for multiple reasons. Discontent with K-12 accountability tests had led many parents to opt their children out of state assessment and states to reduce the time devoted to assessments. Postsecondary institutions, concerned about drops in enrollment and non-diverse student populations, increasingly adopted test-optional policies. The pandemic greatly accelerated the challenges for assessment programs. Tests used for admissions, school accountability, and national and international achievement monitoring were all suspended. What does that result portend for the future of educational assessment? This organized discussion will explore some possibilities.

Session Organizer:

Randy Bennett, ETS

Presenters:

Lorrie Ann Shepard, University of Colorado Boulder
Kristen DiCerbo, Pearson
Paul Nichols, NWEA
James Pellegrino, University of Illinois at Chicago

Applications of Diagnostic Classification Models

9:00 to 10:30 am – Paper Session

Chair:

Fabian Zehner, DIPF | Leibniz Institute for Research and Information in Education, Centre f. Int. Student Assessm.

Participants:

A Longitudinal Diagnostic Classification Model with Polytomous Attributes

Matthew James Madison, University of Georgia; Yu Bao, James Madison University; Seungwon Chung; Junok Kim; Laine Bradshaw, University of Georgia

Longitudinal diagnostic models have been developed to provide estimates of student growth with criterion-referenced interpretations. Different from previous studies employing longitudinal diagnostic models, this study models polytomous attributes. Via simulation, we examine accuracy and reliability. Additionally, we examine different estimation options for cases when full model estimation is not feasible.

A Joint Diagnostic Model for Analyzing Multi-Source Data from Technology-Enhanced Learning Systems

Peida Zhan; Kaiwen Man, UMD; Jonathan Malone, University of Maryland, College Park

This study proposes an innovative diagnostic classification model (DCM) for analyzing multi-source data collected from technology-enhanced learning systems as an extension of the joint DCM for item responses and response times by incorporating visual fixation counts into the model.

An Investigation of Differential Item Functioning in Diagnostic Classification Models

Selay Zor, University of Georgia; Laine Bradshaw, University of Georgia

We investigate differential item functioning (DIF) to ensure test fairness and validity in diagnostic classification models (DCMs). We extend DCM-based DIF detection methods to a general DCM framework and explore the effectiveness via simulation under realistic scenarios. Based on findings, we propose levels for item flagging for data review processes.

Modeling Hierarchical Attribute Structures in Diagnostic Classification Models With Multiple Attempts

Tae Yeon Kwon; Anne Corinne Huggins-Manley, University of Florida; Jonathan Templin, University of Iowa; Mingying Zheng, University of Iowa

The purpose of this study is to develop a sequential HDCM and investigate its impact on classification accuracy in the presence of hierarchies when multiple attempts are allowed in dynamic assessment. This study will therefore provide information to practitioners about possibilities for psychometric modeling of dynamic classroom assessment data.

An Investigation of Longitudinal Diagnostic Classification Using a Hidden Markov Model

Jiajun Xu, University of Georgia; Laine Bradshaw, University of Georgia

This paper investigates longitudinal diagnostic classifications using a hidden Markov model. We apply a newly-adjusted, unconstrained hidden Markov model that allows students to lose and/or regain attributes to better approximate real learning trajectories. We examine the model in over 100+ simulation conditions that strive to mimic practical scenarios.

Discussant:

Benjamin R. Shear, University of Colorado Boulder

NCME Business Meeting and Presidential Address

10:45 to 12:45 pm – Plenary Session

Join your friends and colleagues for the NCME business meeting where an update about our organization will be provided, annual awards will be granted, business and finance report will be shared, and new board members will be introduced. The business meeting will be followed by our traditional Presidential Address.

Assessment Research and Practice in the Post-COVID-19 Era
Ye Tong, Pearson

(Invited Session) Stakeholder Perspectives on Validating Licensure Examinations

1:00 to 2:00 pm - Organized Discussion

This organized discussion will explore multiple perspectives on validity within the context of legal licensure. Staff from the National Conference of Bar Examiners will provide an overview of past, present, and future validity efforts undergirding the current bar examination and the next generation exam. Challenges associated with communicating about validity research with various stakeholder groups will be explored and possible gaps between research and practice will be highlighted, with suggestions offered for potential avenues for bridging those gaps. Discussion will center on practical approaches to prioritizing research activities relating to validity and collaboration across practice and research, with audience participation welcomed.

Session Organizer & Chair:

Joanne Kane, National Conference of Bar Examiners

Presenters:

Kellie Early, National Conference of Bar Examiners
Ken Kraus, National Conference of Bar Examiners
Joanne Kane, National Conference of Bar Examiners
Mark Albanese, National Conference of Bar Examiners

Topics in Test Development

1:00 to 2:00 pm - Research Blitz Session

Chair:

Anna Topczewski, WestEd

Participants:

Confirming Bias in Higher Education: Introducing and Validating a Confirmation Bias Performance Assessment

Gabe Avakian Orona, University of California, Irvine; Remy Pages, University of California, Irvine; Richard Arum, University of California, Irvine; Jacque Eccles, University of California, Irvine

With heightening political tension, a global pandemic, and the threat of fake news, education should produce unbiasedly evaluate evidence and allow themselves to reach uncomfortable conclusions. This proposal—in collaboration with Educational Testing Service (ETS)—introduces a novel confirmation performance assessment, while applying a variety of validation techniques from classical and modern test theory.

Automated task generation for immersive assessments

Martha McCall, McKinsey & Company; Xinchu Zhao, Imbellus; Jeremiah McMillan, Imbellus

Automatic item generation (AIG) systems use item models to create large volumes of items. This study investigates the use of task models for automatically generating immersive simulation task forms. Results of varying selected attributes of generated tasks are discussed.

Using process data to develop indicators for the assessment of group collaboration

Nafisa Awwal, University of Melbourne; Mark Wilson, University of California, Berkeley; Zhonghua Zhang, University of Melbourne

Group collaboration is difficult to assess due to its interactive and ephemeral nature. We use a multidimensional framework for collaborative problem-solving to ground the development of process-based indicators for group collaboration. We unpack development specifics of how these relate to the quality of group CPS with respect to that framework.

Exploring Pretesting Designs for Automatic Generated Items

Fen Fan; Joshua Goodman, NCCPA

There are two goals of this study. First, we explore the matrix sampling designs (Pomplun, 2007) for pretesting AIG items. Second, we seek to establish guidelines under which the Rasch model is robust to potential violations of local item independence arising from the inclusion of item families on a form.

An Experiential approach to test design and validation.

Sergio Araneda, University of Massachusetts Amherst; Stephen Sireci, University of Massachusetts Amherst

We propose an approach to test design and validation based on the study of the experiences of the examinees. We present a conceptual framework to understand the experiences and discuss how this approach will work in practice, for both test design and validation.

Creating Realistic and Informative Simulations with Applications for Psychometric Design Elements

Sukkeun Im, NWEA; Richard Patz, University of California, Berkeley; Melinda Montgomery, NWEA; Christina Schneider, NWEA; Nisha Padminiamma, NWEA

This study examines whether measurement error in theta estimates based on previous year population estimates needs to be adjusted for them to form an appropriate basis for the assumed true theta distribution in a simulation framework. Initial results show evidence that adjustments for measurement error improve simulation quality.

Techniques for Missing Data and Guessing Behavior
1:00 to 2:00 pm - Research Blitz Session

Chair:
Randy Bennett, ETS

Participants:

Evaluating Approaches for Dealing with Omitted Items in Large-Scale Assessments
Seong Eun Hong; Scott Monroe, University of Massachusetts Amherst

Large-scale assessments (LSAs) are low-stakes tests; consequently, examinees might randomly guess or generate no responses. Such disengaged test-taking behavior can undermine the validity of test score interpretation. The present study investigates the impact of how omitted responses are handled on item and person parameter estimates with the ad hoc and model-based approaches.

Exploring the impact of random guessing in distractor analysis
Kuan-Yu Jin, Hong Kong Examinations and Assessment Authority; Wai Lok Siu, Hong Kong Examinations and Assessment Authority; Xiaoting Huang, Hong Kong Examinations and Assessment Authority

Distractor analysis is an important procedure to check the utility of response options within a multiple-choice item. A new IRT model is proposed in this study to detach the influence of random guessing on response option functioning. The mathematics tests of the HKDSE are provided and analyzed for demonstration.

An Application of Explanatory IRT to Predict Factors Contributing to Rapid Guessing
Guher Gorgun, University of Alberta; Seyma Nur Yildirim-Erbasli, University of Alberta; Okan Bulut, University of Alberta

This study examines the factors contributing to rapid guessing in a low-stakes reading assessment. Rapid guesses were analyzed using item- and person-level covariates in explanatory IRT. Results showed that the subtest type is a significant predictor. It is recommended that test developers consider the subtest effect on rapid guessing behavior.

Assumption-free lower bound for Cronbach's alpha when missing data are present
Feng Ji; Heyuan Liu, University of California, Berkeley; Xiaoya Zhang, University of California, Davis

We propose a way to construct lower bounds of Cronbach's alpha when data are arbitrarily missing, which is beyond the common treatments after assuming missing at random (MAR) or missing completely at random (MCAR). We demonstrate its computational effectiveness and practical usage using real data and simulation.

Multiple Imputation Approach to Missingness at the School Level: A Comparison Study
Sinan Yavuz, University of Wisconsin-Madison; Xiaying Zheng, American Institutes for Research; Yifan Bai, American Institutes for Research; Markus Broer, American Institutes for Research

Some NAEP contextual variables have high missingness. When this happens at the school level it creates considerable complications. This study compares two multiple imputation approaches: chained equations (MICE) using R and Blimp packages. The results provide empirical suggestions for researchers dealing with multilevel missing data in large-scale assessments like NAEP.

Topics in Multidimensional Item Response Theory

1:00 to 2:00 pm - Paper Session

Chair:

Jiawei Xiong, University of Georgia

Participants:

Exploring the "Cluster-Independent" Property of Ability Estimators in Bifactor Models

Dandan Liao, Cambium Assessment, Inc.; Frank Rijmen, Cambium Assessment, Inc; Tao Jiang, Cambium Assessment, Inc

The present study explores the property of ability estimators under constrained versions of the bifactor model. It was found that when using the testlet model for both standalone items and items within clusters, the latter do not provide information about student ability if the maximum likelihood estimator is used.

Indeterminacy Issue in Bifactor IRT Modeling

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

The bifactor IRT model has an indeterminacy issue when an item's discriminations on the general and specific dimensions are similar. This issue has only been alluded to (Stone & Zhu, 2015). Through simulations, we provide evidence of such indeterminacy and show how sample size could mitigate any estimation issues.

Estimating School-Level Performance on Test Subdimensions

Briana Hennessy, University of Connecticut; Eric Eric Loken, University of Connecticut

State-wide tests measure overall ability while also providing subscores for specific skill dimensions. The subscores are not sufficiently reliable for individual use, but are sometimes aggregated to provide school level feedback. This paper explores the sensitivity of aggregated subscores to school level general ability effects.

Discussant:

Denis Dumas, University of Denver

Involve me and I learn: Applying culturally responsive assessment practices to equitably measure learning of Indigenous students in North America

1:00 to 2:00 pm - Coordinated Paper Session

In the invited panel session that was held virtually on August 31, 2020 (https://www.youtube.com/watch?v=TiUZamOVzjE), discussion focused on addressing the unique challenges Indigenous K-12 students in North America face within the context of traditional systems of learning and assessment. The key question driving the discussion was: How can we help indigenous students succeed in an educational system that has failed them? In a desire to continue dialogue on this important topic the NCME Diversity Issues in Testing Committee is pleased to offer a coordinated session at the NCME 2021 virtual conference focused on showcasing the latest research on culturally-responsive assessment practices for Indigenous students. Traditional forms of assessments are often ineffective and even destructive for Indigenous students (Trumbull & Nelson-Barber, 2019). Culturally-responsive assessment is a promising approach towards fair and equitable assessment but there is a need for more research focused on the Indigenous context and some practical guidance from experts. Therefore, in this session four researchers from the United States and Canada will present their research findings on culturally-responsive assessments that have been proven to be effective in assessing the learning of Indigenous students in a more fair and equitable way.

Session Organizer:

Raman Grover, BC Ministry of Education

Participants:

What Indigenous Students Know and Can Do: Towards Culturally-Responsive and Learner-Relevant Assessments

Madhabi Chatterji, Teachers College, Columbia University

Culturally Responsive Pedagogy: Indigenizing Curriculum

Karen Ragoonaden, University of British Columbia
Kerry Englert kenglert@senecaconsulting.org

Critical Intersections: Engaging and Listening to Native Students' Insights on Instrument Content

Kerry Englert, Seneca Consulting, LLC
Pohai Shultz pohai@hawaii.edu Presenter

Opportunity for who? The potential of opportunity to learn data for culturally and linguistically diverse students

Pohai Kukea Shultz, University of Hawaii at Manoa

Discussant:

Mandy Smoker Broaddus, Education Northwest

Graduate Student Electronic Board Session

1:00 to 2:00 pm

Participants:

Account for Rater Effects with Bayesian Rater Model in Disjoint Rating Design

Xinyue Li, Penn State University

Proper data linkage for rater severity estimation is very important for sparse rating design. In particular, researchers have not yet considered how various characteristics of linking sets impact the effectiveness of estimating true ability. The current study examined the performance of a proposed Bayesian IRT rater model through simulation.

A CDM Approach to Explore the Validity of Number Puzzles of CogAT7

Qingzhou Shi; Joni M. Lakin, University of Alabama, Wenchao Ma, University of Alabama

This study will investigate the validity of the Number Puzzles in the Quantitative battery of the Cognitive Abilities Test, Form 7 (CogAT7) using cognitive diagnostic modeling (CDM) and the CogAT7 national standardization data (N=65,630), with the ultimate goal of extending CDM to all three batteries and nine subtests.

A Joint Modeling of Response Accuracy and Time with Automatically Generated Items

Yan Yan, Georgia Tech; Susan Embretson, Georgia Institute of Technology

The hierarchical framework for modeling accuracy and response time (vin der Lindon, 2007) provides a promising approach to simultaneously explore the two important sources of information about test takers. The study here applies this model with the automatically generated items, in the meantime assesses the impact of cognitive process modeling.

A Motivational-Developmental Free Response Assessment through a Testlet Lens

David Alpizar; Brian French, Washington State University

A four-item writing prompt assessment measures university student's motivational and developmental attributes. This assessment format may violate local independence. This study examined the local dependency and theoretical-testlet models for the assessment. Local item dependence was present. A scoring inference with a reduced testlet model for the assessment was supported.

Analysis of Digital Reading Processes from Multimodal Time-series Data Using Deep Learning

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

In this study we designed and trained a recurrent neural network (RNN) to analyze multimodal time-series reading process data. Preliminary results showed that including the time-series process data through the use of the RNN resulted in an increase in accuracy when predicting students' performance on items.

Applying Latent Variable Approach for Examining Measurement and Prediction Invariances

Tuba Gezer

The purpose of this study is to examine measurement invariance (MI), and prediction invariance (PI) simultaneously based on latent scores and compare results with observed scores using moderated multiple regression analysis (MMR) with ordinary least square estimation and two-stage least squares (2SLS) estimators using English Language Learner high-stakes testing results.

Applying Model Regularization in Identifying Predictors of Student Achievement

Mingqin Zhang, University of Iowa; Jihong Zhang, University of Iowa; Guanlan Xu

Few studies have utilized the full range of variables provided by PISA when finding effective predictors for student performance. This study applies model regularization techniques (such as Ridge, LASSO, and Elastic Net) on PISA 2015 and aims to identify important predictors for U.S. student achievement.

Comparison of Procedures for Detecting Drifted Items under 3PL and Rasch Models

Kuo-Feng Chang; Won-Chan Lee, University of Iowa

The relative performance of each item parameter drift detection method under the 3PL model may not necessarily lead to the same conclusion under the Rasch model. In this study, we examined the effectiveness of a few commonly used detection methods under both 3PL and Rasch models using simulated data.

Consequences of assuming: Effects on bias and efficiency of IRTree trait estimates

Nikole Gregg, Cambium Assessment, Inc.; Brian C Leventhal, James Madison University

IRTree models, along with other multidimensional item response theory models, assume an underlying response process. We investigate bias and efficiency of substantive trait estimates from models correctly assuming the response process, and models incorrectly assuming the response process. Results inform the importance of conscientiously selecting the appropriate response process model.

Detecting Gender Difference in PISA Mathematics Anxiety Items for Canada and Finland

Lindsay Coppens, Ontario Institution for Studies in Education

Research has explored gender differences in mathematics anxiety and performance in large-scale assessments; however, it has not examined the possibility of different interpretations of mathematics anxiety items. This paper uses DIF and DSF to detect gender differences in the interpretation of PISA's 2012 mathematics anxiety items for Canada and Finland.

Development of a Person-fit Statistic for Dynamic Measurement Modeling

Yixiao Dong, University of Denver; Denis Dumas, University of Denver

The study aims to formulate and evaluate a person-fit statistic for Dynamic Measurement Modeling (DMM). A residual-based person-fit formula in IRT was adapted and developed into the person-fit statistic for DMM. A follow-up simulation study was also planned to demonstrate the effectiveness of this index.

Equating Errors of Applying Unidimensional IRT Methods to Equate Multidimensional Test Forms

Haimiao Yuan, University of Iowa

This study explores equating errors when applying unidimensional IRT equating methods to multidimensional test forms. The effects of dimensional structure consistency in different forms are discussed. We also explore how large the correlation between dimensions needs to be to achieve acceptable equating results using unidimensional IRT methods.

Examining English Language Learners' Proficiency in terms of Gaps

Tuba Gezer; Brian Gong, Center for Assessment

Reducing achievement gaps is necessary to increase the quality of education for every student. This study examines the relationships between English Learners' achievement and EL density, the relationship between ELs' school achievement status and school growth and discusses the reachability of state goals for ELs at school and districts levels.

Generalized Graded Unfolding Model-Asymmetric: A Modified GGUM for Asymmetric Attitude Data

Emily Chai; James S. Roberts, Georgia Institute of Technology

We investigate the possibilities to expand the original Generalized Graded Unfolding Model (GGUM, Roberts et al., 2000) by fitting asymmetric data. Responses to abortion attitude questionnaires are analyzed with both GGUM and GGUM-Asymmetric to compare model fit. Parameter recovery is explored with simulated data.

Handling Missing Non-Normal Ordered Categorical Data

Hacer Karamese; Juliana Cerentini Pacico, University of Iowa

The performance of predictive mean matching for missing data in non-normally distributed ordinal data were compared to original data, listwise deletion, and pairwise deletion using a Monte Carlo simulation. In general, the predictive mean matching method fell behind the other methods under consideration.

Handling Outliers in Random Coefficient Multilevel Model Using Two Alternative Approaches

Jia Quan; Walter Leite, University of Florida; Yuxi Qiu, Florida International University

This project intends to compare the performance of two newer alternative approaches, the heavy-tailed method, and the rank-based method when fitting a random coefficient multilevel model under various conditions through a simulation study. The authors intend to provide some evidence in using these alternative approaches for applied researchers to consider.

Identifying Careless Responding and Response Patterns for Negatively Keyed Items

Jacqueline King, James Madison University

This study investigates methods of identifying careless responding, specifically emphasizing the inclusion of negatively keyed items. Results are used to broadly examine 1) the efficacy of using negatively keyed items to identify careless responses (as compared to other methods), and 2) the cognitive difficulties of responding to negatively keyed items.

If Differential Distractor Functioning Occurs, Must Differential Item Functioning Occur?

Jiayi Deng

This study aimed to examine the relationship between differential distractor functioning (DDF) and differential item functioning (DIF) in multiple-choice items from the PIRLS achievement test. Generalized linear models were utilized for DDF and DIF detection, and their relationship were examined via correlation and binomial tests.

Mapping both items and persons using simultaneous GGUM and MSI

Na Liu; James S. Roberts, Georgia Institute of Technology

Both favorability judgments to scale item locations with the MSI models and graded disagree-agree responses to estimate item (and persons) locations with the GGUM contain theoretically similar information about item locations on the latent continuum. Using both types of data in an IRT model can get a more precise estimates.

Model-data Fit Evaluation of the Joint Model of Responses and Response Times

Xin Qiao

The current study evaluates the model-data fit of the joint model of item responses and response times using the post-data simulation method when maximum likelihood estimation is used. The results indicate that the post-data simulation method is a promising tool in examining of the model-data fit of the joint model.

Modeling Latent Differential Rater Functioning Drift Using Signal Detection Theory

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

This study proposes a longitudinal mixture framework using latent class signal detection theory to detect differential rater drift among different latent subgroups. Real-world data analysis identified individual variations in changes of rater behavior over time in the context of latent subgroups. Simulation studies provide inferences on estimation and parameter recovery.

New Generalized Graded Unfolding Model with Time Series Component

Zhaoyu Wang, Georgia Institute of Technology; James S. Roberts, Georgia Institute of Technology

This study will apply the GGUM (Generalized Graded Unfolding Model) to estimate people's attitudes toward gun control policy and ARIMA model to explore the correlation between people's attitudes and mass shootings from 2009 to 2013. This research will include time series factor into the GGUM to develop a new model.

Regularization in G-DINA model in Cognitive Diagnosis

Yuan Ge, University of Alabama

Cognitive diagnostic analysis (CDA) provides finer grained information. However, the complexity of current cognitive diagnostic models posts a problem. This study aims to add a regularized parameter to the G-DINA model to simplify the model and evaluate the regularization performance, while keeping the classification accuracy of the regularized models.

Students' Perceptions of (un)Fairness in PISA Data: Assessing Cross-cultural Measurement Invariance

Amir Rasooli; Amin Mousavi, University of Saskatchewan

Empirical research has largely investigated students' perceived classroom fairness within a cultural context. To expand this research cross-culturally, this study analyzed the measurement invariance of PISA 2015 survey on students' perceived unfairness. The results showed a lack of metric invariance, implying that the participants' unfairness perceptions vary cross-culturally.

The Impact of Six Missing Data Handling Methods on Scale Linking Accuracy

Tong Wu; Stella Kim, University of North Carolina at Charlotte; Carl Westine, University of North Carolina at Charlotte

This purpose of the study is to evaluate the impact of six missing data handling approaches on IRT scale linking accuracy. Under various simulation conditions, the relative performance of the missing data handling methods on both Stocking-Lord and Haebara approaches is explored to inform the practitioner the most precise and accurate approach.

The Level-Specific Fit Evaluation in MCFA with Different Factor Structures across Levels

Bitna Lee, Kyungpook National University; Sohn Wonsook, Kyungpook National University

A Monte Carlo study was conducted to investigate the performance of level-specific fit indices derived by a partially saturated model method in Multi-level Confirmatory Factor analysis. This study extended previous studies by examining their performance under (a) MCFAs with different factor structures across levels and (b) more various design factors.

Towards Automated Essay Scoring and Feedback Generation

Chang Lu, University of Alberta; Maria Cutumisu, University of Alberta; Mark Gierl, University of Alberta

We implemented a unified model for automated essay scoring and feedback generation employing a word-embedding technique, a deep learning model, and a constrained Metropolis-Hastings sampling using the Markov Chain Monte Carlo method. The model yields high performance on evaluating essays across domains and generates fluent and coherent feedback sentences.

Understanding Problem-Solving Styles in Technology-Rich Environments by Log Data Analysis

Yizhu Gao; Xiaoming Zhai, University of Georgia; Okan Bulut, University of Alberta; Ying Cui, University of Alberta

To examine and compare problem-solving styles in technology-rich environments (TRE), we abstracted two behavioral indicators (planning duration and interaction frequency) from log data of problem solving in TRE (PSTRE). Results confirmed the presence of Acting/Reflecting styles together with the Shirking style and the superiority of Acting style in PSTRE.

Understanding Relationships Between a College Tracker Program and Student Post-Secondary Plans

Catherina Villafuerte, University of Connecticut

The purpose of the study is to examine implementation of a matching and tracking tool, a low-cost college application intervention, in a large urban school district in the northeastern US. Additionally, the relationship between various implementation metrics and student post-secondary plans will be explored.

Generalizability Theory Applications

1:00 to 2:00 pm - Paper Session

Chair:

Stella Kim, University of North Carolina at Charlotte

Participants:

Integrating Parallel Splits into Generalizability Theory Analyses

Walter Vispoel, University of Iowa; Guanlan Xu; Wei Schneider

Using items as tasks within multi-facet generalizability theory designs will typically produce conservative estimates of reliability estimates that reflect random rather than classical parallelism. We demonstrate how properly balanced splits can address these issues by improving score consistency, reducing measurement error, and producing reliability indices that approximate classical parallelism.

Applying Multivariate Generalizability Theory to Automated Essay Scoring for English Language Learners

Dandan Chen, The American Board of Anesthesiology; Joshua Wilson, University of Delaware

We applied multivariate generalizability theory to evaluate the reliability and validity of English Language Learners (ELLs) and non-ELLs' writing scores produced by the PEG Writing® automated essay scoring system. The results showed different contributions of facets to measurement error and discrepancy in coefficients for the two groups.

Linkages Between Latent State-Trait and Generalizability Theories

Walter Vispoel, University of Iowa; Wei Schneider; Guanlan Xu

Despite fundamental differences in focus, latent state-trait theory and generalizability theory approaches to understanding psychometric properties of scores share much in common. We use a structural equation modeling framework to illustrate similarities and differences between the theories and demonstrate how the same data can be readily interpreted from both perspectives.

Discussant:

Ji Seung Yang, University of Maryland

Leveraging Process Information in International Large-Scale Assessments: Recent Findings from PIAAC

2:15 to 3:45 pm - Coordinated Paper Session

This symposium highlights advanced psychometrics used in four studies to address questions on how process information such as timing data and sequences of actions are related to task performance and how to use such information to interpret test takers' achievements and identify variations among groups/countries in large-scale assessments. Process data collected in the Programme for the International Assessment of Adult Competencies (PIAAC) are used as illustrative examples in this coordinated session. The first paper leverages timing data to investigate the relationship between the willingness of individuals to engage with cognitive assessments in relation to item position and variations in item difficulty. The second paper examines whether process-based information such as problem-solving strategies indicated by action sequences could better explain differential item functioning (DIF) by latent classes given the same ability level. The third paper focuses on using timing and navigation information to identify and interpret age effects in dealing with information from search-engine environments. The fourth paper assesses the consistency of test-taking behaviors across multiple items by using both aggregate-level response process variables and action sequences. These studies show the promise of leveraging process information to improve proficiency estimation and the validity of test score interpretations in large-scale assessments.

Session Organizer:

Qiwei He, Educational Testing Service

Chairs:

Qiwei He, Educational Testing Service

Frank Goldhammer, DIPF | Leibniz Institute for Research and Information in Education, Centre f. Int. Student Assessm

Participants:

Willingness to Engage with a Low-Stakes Assessment: Evidence from a Natural Experiment in PIAAC

Francesca Borgonovi, University College London; Organisation for Economic Co-operation and Development; Francois Keslair, OECD; Marco Paccagnella, OECD

Exploring Group Differences in Large-Scale Assessments Using Latent Class Analysis on Process Data

Daniella Rebouças-Ju, University of Notre Dame; Qiwei He, Educational Testing Service; Xiang Liu, Educational Testing Service

Effects of Age in Dealing with Information from Search-Engine Environments: Results from an Analysis of PIAAC Log File Data

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, Centre f. Int. Student Assessm; Ulf Kroehne

Evaluating Consistency of Behavioral Patterns across Multiple Tasks Using Process Data in PIAAC

Qiwei He, Educational Testing Service; Dandan Liao, Cambium Assessment, Inc.; Hok Kan Ling, Queen's University; Hong Jiao, University of Maryland

Discussant:

Matthias von Davier, Boston College

(Invited Session) Assessments For Different Purposes: Issues on Scoring, Score Use, and Measurement

2:15 to 3:45 pm - Organized Discussion

Assessment is an important part of ensuring equitable access to education and professional opportunities. It also plays a critical role in learning. Not only do learners demonstrate their knowledge through assessments, test-taking has been shown to enhance learners' understanding of the materials. In recent years, assessment practices have evolved in response to the changing landscape of education and workforce. The COVID-19 pandemic of the past year has accelerated these changes. As measurement professionals, how do we ensure the assessments we develop continue to serve stakeholders? How do we leverage technological advances to capture richer information with our products? How can we ensure assessments serve their intended purposes? Join R&D leaders in education, workforce development, and professional credentialing to learn about the unique challenges and opportunities in these industry sectors. Speakers will share research from their programs, as well as discuss issues surrounding measurement and score use.

Session Organizer & Chair:

Ada Woo, Ascend Learning

Presenters:

Christine Mills, Ascend Learning
Ou Lydia Liu, ETS
Michelle Derbenwick Barrett, Edmentum
Carol Ezzelle, National Board for Professional Teaching Standards

The Impact of COVID-19 on Educational Measurement, Part 1: K-12 Assessment

2:15 to 3:45 pm - Coordinated Paper Session

In the summer of 2020, Educational Measurement: Issues and Practice editor Deborah Harris invited selected testing and measurement experts to assess the impact of the COVID-19 pandemic on the field for a special issue of the journal. In two separate symposia, contributing authors to this special issue will present their evaluation of the impact of COVID-19 on the profession, including their best ideas about how the educational measurement community can and should respond to these unprecedented challenges. This first symposium includes papers related to K-12 assessment. The second symposium includes papers related to admissions and certification testing. Presentations will be short, emphasize cross-cutting themes, and leave plentiful time for audience questions. Into the Unknown: Assessments in Spring 2021 Leslie Keng, Michelle Boyer, Scott Marion What Hath the Coronavirus Brought to Assessment? Unprecedented Challenges in Educational Assessment in 2020 and Years to Come Hong Jiao, Robert Lissitz Remotely Proctored K-12 High Stakes Standardized Testing During COVID-19: Will it Last? Rochelle Michel How Can Released State Test Items Support Interim Assessment Purposes in an Educational Crisis? Emma Klugman, Andrew Ho Working with atypical samples Zhongmin Cui Educational Assessment of the Post-Pandemic Age: Chinese Experiences and Trends Based on Large-Scale Online Learning Hong Su

Chair:

Deborah Harris, University of Iowa

Participants:

Into the Unknown: Assessments in Spring 2021

Leslie Keng, Center for Assessment; Michelle Boyer, Center for Assessment; Scott Marion, Center for Assessment

What Hath the Coronavirus Brought to Assessment? Unprecedented Challenges in Educational Assessment in 2020 and Years to Come

Hong Jiao, University of Maryland; Robert Lissitz, University of Maryland

Remotely Proctored K-12 High Stakes Standardized Testing During COVID-19: Will it Last?

Rochelle Michel, Curriculum Associates

How Can Released State Test Items Support Interim Assessment Purposes in an Educational Crisis?

Emma M. Klugman, Harvard Graduate School of Education; Andrew Ho, Harvard Graduate School of Education

Working with atypical samples

Zhongmin Cui, CFA Institute

Educational Assessment of the Post-Pandemic Age: Chinese Experiences and Trends Based on Large-Scale Online Learning

Hong Su, China National Institute of Education Sciences

Discussant:

Andrew Ho, Harvard Graduate School of Education

Developing an Alternate English Language Proficiency Assessment within a Principled Design Framework

2:15 to 3:45 pm - Coordinated Paper Session

The development of an alternate assessment for students with the most significant cognitive disabilities is challenging; more so when the assessment is also a measure of English language proficiency (ELP). The challenges are clear:

- There is no agreed-upon definition of the small, diverse student population of English Learners with the most significant cognitive disabilities (ELSCDs).
- The construct (ELP) may manifest for this student population (ELSCDSs) in ways not typically observed in general assessments, making the assessment extremely complex.
- Few SMEs exist for ELSCDs
- There are no exemplar PLDs to support this unique construct and population

Thus, the linear development of assessment components under a principled assessment design framework may result in less than optimal results than for less complex constructs. The five presenters describe a coordinated effort to address these challenges. Presenters supporting the Collaborative for the Alternate Assessment of English Language Proficiency (CAAELP) will discuss:

- a principled, explicitly-iterative validity framework guiding assessment system coherence,
- an overview of relevant assessment design features and considerations,
- the iterative development of PLDs,
- item development and alignment strategies, issues, and considerations, and
- the psychometrics models and reporting metrics for this relatively new assessment program.

Session Organizer:

Daniel Lewis, Creative Measurement Solutions LLC

Chair:

Edynn Sato, Sato Education Consulting LLC

Participants:

A Principled, Explicitly-Iterative Validity Framework in Support of Assessment System Coherence
Daniel Lewis, Creative Measurement Solutions LLC

Valid Assessment of English Learners with Significant Cognitive Disabilities: Design and Considerations
Edynn Sato, Sato Education Consulting LLC

Iterative Development of PLDs for an Alternate English Language Proficiency Assessment
Nami Shin, UCLA CRESST

Principled Item Development and PLDs for English Learners with Significant Cognitive Disabilities
Kelly Ickes, Cognia; David Sanderson, Cognia; Steve Ferrara, Cognia

Psychometric Perspectives of Developing an Alternate English Language Proficiency Assessment
Nami Shin, UCLA CRESST; Li Cai, UCLA

Discussant:

Audra Ahumada, Arizona Department of Education

Topics in Validity

2:15 to 3:45 pm - Paper Session

Chair:

Yanyan Fu, GMAC

Participants:

Graduate Record Examination Predictive Validity: A Systematic Review of Empirical Research

Maureen Font, University of Illinois - Chicago; Yue Yin, University of Illinois - Chicago

Using Graduate Record Examination (GRE) scores in graduate admissions is controversial. Programs need to evaluate the usefulness of the GRE, when facing the trend of dropping GRE scores and the limitation of other predictors. We synthesize GRE predictive validity studies since 2012 and shed light on this critical issue.

Using Automated Feedback to Develop Writing Proficiency

Katherine Huang; Joshua Wilson, University of Delaware

This study examined 431 fifth and fourth graders' growth in writing using an automated writing evaluation (AWE) software system and whether its prolonged usage had any unaided/aided transfer effects. Results revealed a logarithmic shape of growth and showed that there was no unaided transfer effect, but an aided transfer effect.

Characterization of Written Feedback in the Context of Virtual Formative Assessment

Sandra Cecilia Zepeda, Universidad Catolica; Valeria Carolina Zunino, UC Davis

This study examines the characteristics of teachers' written feedback in the context of virtual classes. Evidence obtained from one school in Chile was analyzed using a categorical analysis matrix. Results show that their characteristics are not aligned with the formative assessment approach reflecting the need to support teachers' assessment literacy.

Building Bridges During a Pandemic: Community Validity as Foundational to Score Credibility

Kerry Englert, Seneca Consulting, LLC; Pohai Kukea Shultz, University of Hawaii at Manoa; Karla Egan, EdMetric, LLC

Often, large-scale assessments are seen as detached from daily classroom instruction, and this is especially true during a pandemic. In order to build relevancy, the Kaiapuni Assessment of Educational Outcomes has intentionally focused on deep collaborations with the community and the notion of community validity to guide assessment development.

In-Person Proctoring Versus Remote Proctoring with a Medical Licensing Examination

Maxim Morin; Andre De Champlain, AstraZeneca; Cecilia Alves, Medical Council of Canada

With the advent of COVID-19 pandemic, remote proctoring has emerged as a promising alternative for delivering examinations. This study evaluates and compares the impact of technical issues on test scores in a high-stake medical licensing examination between in-person proctoring and remote proctoring.

Discussant:

Stanley N Rabinowitz, Pearson

Differential Item Functioning (DIF) Applications

2:15 to 3:45 pm - Paper Session

Chair:

Danielle Guzman-Orth, Educational Testing Service

Participants:

Investigating Assessment Conditions Potentially Associated with DIF

Qiao Lin, University of Illinois at Chicago; Jeffrey Steedle, ACT

This study investigated associations between DIF and the properties of items and examinees. In empirical analyses, easier items were more likely to exhibit DIF favoring females. Simulation studies indicated that difficulty, discrimination, test length, examinee ability, and sample size were not associated with inflated Type-I error for DIF flags.

Using Mantel- Haenszel Procedure to Assess Differential Bundle Functioning: A Meta-Analysis Approach

Lanrong Li; Betsy Jane Becker, Florida State University

We propose using meta-analysis techniques to synthesize differential item functioning (DIF) and assess differential bundle functioning (DBF). The Mantel- Haenszel procedure was used to illustrate the approach. We found that with correlated dimensions underlying DIF items, the proposed DBF test performed well in Type-I error and power rates.

Dissecting Ability in DIF Analysis

Rabia Esma Sipahi, University of Kansas; John Poggio, University of Kansas

Measuring true ability is paramount to DIF analyses. We propose measuring ability using examinee skill or indicator scores (rather than as customary, the assessment total score) and explore generalizability for impacted groups. Using NAEP 2015 assessments, we are discovering that defining true ability more exactingly yields more accurate race, gender, and ELL rooted DIF results.

IRT Residual Approach to Detecting Differential Item Functioning

Hwanggyu Lim, Graduate Management Admission Council; Edison M. Choe, Graduate Management Admission Council; Kyung (Chris) T. Han, Graduate Management Admission Council; Sung-Hyuck Lee; Minju Hong, University of Georgia

Among the plethora of DIF detection techniques in literature, we introduce a new IRT residual approach that particularly stands out for its simplicity without sacrificing efficacy. A preliminary study demonstrates its incredible potential as a quick and surprisingly powerful method that outperforms the competition, even with relatively small sample sizes.

An Analysis of DIF and Sources of DIF in Motivation Items *Jacquelyn A.*

Bialo; Hongli Li, Georgia State University

This study used pairwise comparisons and multiple-group DIF with a base group to evaluate differential item functioning in PISA 2015 achievement motivation items across gender and ethnicity before and after using anchoring vignettes to account for the impact of group differences in response scale use as a source of DIF.

Discussant:

Nina Deng, Kaplan INC.

Applications in Adaptive Testing

2:15 to 3:45 pm - Paper Session

Chair:

Kevin Krost, Fralin Biomedical Research Institute

Participants:

Investigating Hybrid Test Designs in Passage-based Adaptive Tests

Ye Ma, University of Iowa; Deborah J Harris, University of Iowa; Stephen B. Dunbar, University of Iowa

The proposed HMCAT designs fill the gap of current literature and practice on hybrid designs and passage-based adaptive testing. Findings support the efficiency and flexibility of HMCAT designs from a practical perspective. This study provides implications for practitioners on how to decide, evaluate and maintain these designs effectively in practice.

An Adaptive Testing Procedure to Inform Transitions from High School to College

G. Gage Kingsbury

High school students struggle to identify reasonable paths to college. While resources are available, students often see an incomplete picture. The current study explicates a composite adaptive testing procedure that assesses student interests and achievement to provide information about pathways to college, to help students evaluate and expand their options.

Investigating the Need for Cognitive Level Constraints in a Computer Adaptive Assessment

Melinda Montgomery, NWEA; Christina Schneider, NWEA; Sukkeun Im, NWEA

We investigate if cognitive level (DOK, Webb, 2005) constraints for an ELA CAT are necessary. A commonly implemented requirement carried over from fixed-form assessments is that a CAT should constrain a particular proportion of items on cognitive complexity. We find this may not be needed.

Comparative Study of Three Mixed-Format Adaptive Test Designs

Hsin-Ro Wei, Riverside Insights; Unhee Ju, Riverside Insights; JongPil Kim, Riverside Insights

Although many studies with CAT have been conducted, there are a few studies that investigated mixed-format CAT designs with both discrete items and testlets. This study is designed to compare the measurement properties of ability estimates across the three mixed-format designs proposed under the full and hybrid CAT frameworks.

Differential Item Functioning in Multistage Testing

Ru Lu, Educational Testing Service; Paul Adrian Jewsbury, Educational Testing Service

This study investigates the impact of matching criterion of the Mantel-Haenszel statistics on the accuracy of differential item functioning detection with multistage testing data. The matching approaches include different raw scores (block, book, or pooled book) and equated scores. Finally, the observed DIF measures are compared with IRT-based DIF measures.

Discussant:

Laurie Davis, Curriculum Associates

(Invited Session) Pivoting in a Pandemic

4:00 to 5:30 pm - Organized Discussion

Impacts of the pandemic have forced assessment methodologies to be adapted. In this session, participants will discuss challenges they faced, research they drew from, and their solutions. Iowa Testing Programs will discuss addressing potential pandemic-related sensitivity concerns for large-scale assessment. This presentation will describe a nontraditional, technology-driven methodology created to identify, review, and evaluate potential areas of concern across the full item bank of a large-scale assessment program. HumRRO will discuss approaches, as well as benefits and challenges, associated with conducting assessment reviews virtually in an asynchronous setting. They will describe two recently conducted activities completed in this manner—an item alignment and standard alignment. The National Board for Professional Teaching Standards will provide an overview of how it expanded its scoring model from regional to synchronous. This presentation will describe operational considerations made to expand the synchronous scoring model that led to successful scoring and continuity of the assessment cycle, despite the pandemic. Massachusetts Comprehensive Assessment System staff will present a punch list of operational and research studies being conducted to generate accurate achievement and growth results and to establish stable trends in 2021. Results on student demographics and the representativeness of the 2021 MCAS examinees will be presented.

Session Organizer:
Jennifer Beimers, Pearson

Chair:
Trent Workman, Pearson

Presenters:
Tim Hazen, Iowa Testing Programs
Emily Dickinson, HumRRO
Yvette Nemeth, HumRRO
Andrea Hajek, National Board of Professional Teaching Standards
Bob Lee, Massachusetts Department of Education
Kathleen Flanagan, Massachusetts Department of Education

Probabilistic Graphical Models for Writing Process Data

4:00 to 5:30 pm - Coordinated Paper Session

Most of the research literature around writing process, in the field of educational measurement, has primarily focused on feature development and exploring the applications of the writing process features for the purpose of essay scoring, gaining new knowledge about writing and different writing tasks, understanding individual or subgroup differences, improving test validity, as well as profiling of students' writing patterns or writing styles. Less work has targeted at statistical modeling of the writing features or the overall writing process. In this symposium, we present four papers around using probabilistic graphical models to describe and understand writers' text-production process. Despite of its increasing popularity in the psychometrics literature, most development in graphical modeling has been centered on the modeling item response; its application in treating timing and process data is much rarer. We hope this symposium will not only demonstrate the utility of this type of models in treating writing process data, but also inspire greater interests in this line of research among the measurement community when dealing with timing and process data.

Session Organizers:

Hongwen Guo, Educational Testing Service
Xiang Liu, Educational Testing Service

Chair:

Hongwen Guo, Educational Testing Service

Participants:

Effects of Scenario-Based Assessment on Students' Writing Processes

Hongwen Guo, Educational Testing Service; Mo Zhang, Educational Testing Service; Paul Deane, ETS; Randy Bennett, ETS

Examine the Prompt Effects on Composition Process Using Mixed Markov Process Model

Mo Zhang, Educational Testing Service; Xiang Liu, Educational Testing Service; Hongwen Guo, Educational Testing Service

Consistency and Predictive Power of Keystroke Feature Variables over Time

Mengxiao Zhu, Educational Testing Service; Xiang Liu, Educational Testing Service

Bayesian Nonparametric Ordered Discrete Latent Variable Models

Xiang Liu, Educational Testing Service

Discussant:

Matthew Johnson, ETS

Mode Comparability in College Admissions Testing: In-depth Investigations and Methodological Considerations

4:00 to 5:30 pm - Coordinated Paper Session

With rapid advances in technology, testing programs are increasingly shifting from paper-based to computer-based modes of administration. This session focuses on mode comparability in the context of college admissions testing. The first presentation provides context through a review of mode comparability research from 2010–2020 and a study of trends in mode comparability over time. The three subsequent presentations offer a detailed study of mode comparability for the ACT assessment based on three experiments conducted during the 2019–2020 academic year. This includes estimated mode effects in five content areas (English, math, reading, science, and writing), an investigation of whether mode effects are similar across demographic groups and for examinees of varying ability, and an examination of evidence that observed mode effects might be related to differences in speededness between paper-based and computer-based testing. The final presentation addresses the validity of non-experimental matching methods for evaluating mode effects. From this session, attendees should gain a deeper understanding of mode comparability, including potential causes of mode effects, methods of investigation, and deciding whether an adjustment is needed to support score comparability.

Session Organizer:

Jeffrey Steedle, ACT

Participants:

Paper and Online Testing Mode Comparability: A Literature Review from 2010–2020

Ann Arthur, ACT; Shalini Kapoor, ACT

Three Studies of Comparability Between Paper-Based and Computer-Based Testing for the ACT

Jeffrey Steedle, ACT

Investigation of Differential Mode Effects When Comparing Paper-Based and Computer-Based ACT Testing

Lu Wang, ACT; Jeffrey Steedle, ACT

Speededness as a Possible Explanation for Mode Effects on the ACT

Shichao Wang, ACT; Dongmei Li, ACT

Comparing Non-Experimental Methods of Evaluating Mode Comparability

YoungWoo Cho, ACT; Xin Li, ACT

Discussant:

Laurie Davis, Curriculum Associates

An Assessment Development and Management (ADM) System for Educational Applications

4:00 to 5:30 pm - Coordinated Paper Session

This symposium introduces an assessment development and management (ADM) system for educational applications called the BEAR Assessment System Software (BASS). The software is based on the BEAR Assessment System, a measurement approach that highlights robust links to cognitive modelling using the construct mapping approach. BASS capitalizes on this framework to offer tools for construct development, item design and development, outcome space design, test design and delivery, test scoring, calibration using Rasch models (both unidimensional and multidimensional), data analysis tools for reliability and validity investigations (including internal structure analyses, and relations with other variables) and a graphical reporting suite that makes abundant uses of the communicative strengths of Wright maps. The first presentation gives an overview of the philosophy and structure of the software system. The second presentation focusses on uses of the system in a classroom setting. The third paper explores the possibilities for adding extra tools and features to the system, using two specific examples: the incorporation of an animation app into item development, and the addition of a verbal item delivery option. The final paper explains how the system can be used in a theory-based educational approach to learning measurement principles through assessment development.

Session Organizer:

Mark Wilson, University of California, Berkeley

Chair:

David Torres Irribarra, Pontificia Universidad Católica de Chile

Participants:

System Overview of an ADM

David Torres Irribarra, Pontificia Universidad Católica de Chile; Mark Wilson, University of California, Berkeley

Classroom Uses of an ADM

Linda Morell; Sara Dozier, Stanford University; Perman Gochyyev, University of California, Berkeley

Adapting New Features into an ADM

Perman Gochyyev, University of California, Berkeley; David Torres Irribarra, Pontificia Universidad Católica de Chile

Teaching Measurement to Graduate Students using an ADM

Mark Wilson, University of California, Berkeley; Aubrey C Condor

Discussants:

Richard Patz, University of California, Berkeley

Derek Briggs, University of Colorado

Multistage Testing with Multiple Subscales: An Investigation of Design and Analysis

4:00 to 5:30 pm - Coordinated Paper Session

In multistage testing (MST), when the routing decision is based on the combined ability estimates from multiple subscales and the collected data are then fitted with multiple unidimensional IRT models, the missingness in the IRT model is not at random, which leads to biased item estimates. This coordinated session presents three potential solutions to the problem from both MST design and data analysis perspectives. The first paper approaches the problem from an MST assembly design and administration perspective. It explores the feasibility of applying a subscale-level MST design, where the routing decisions are made at subscale levels rather than at the combined ability estimates from multiple subscales, to remedy the problem at hand. The second paper investigates via simulations whether using a multiple-imputations approach to address this problem is feasible. Various imputation methods are carried out via multiple imputations by chained equations. The performance of the methods is assessed by comparing item estimates against true values. The third paper explores whether modeling routing block items from other subscales as auxiliary variables using multiple unidimensional IRT scaling could yield unbiased item parameters. The results could provide insights on how to alleviate biased estimation issue in calibrating MST data with many subscales.

Session Organizer:

Xiaying Zheng, American Institutes for Research

Chair:

Markus Broer, American Institutes for Research

Participants:

MST Assembly in Large Scale Assessments: A Top-Down Approach Application

Tong Wu; Young Yee Kim, American Institutes for Research; Xiaying Zheng, American Institutes for Research

Using Multiple Imputations to Handle Non-Random Missing in Multidimensional Multistage Testing

Sinan Yavuz, University of Wisconsin-Madison; Xiaying Zheng, American Institutes for Research; Young Yee Kim, American Institutes for Research

Is MIRT Scaling Necessary for MST with Multiple Subscales?

Xiaying Zheng, American Institutes for Research; Young Yee Kim, American Institutes for Research

Discussant:

Mark Reckase, Psychometric Solutions

Impact of Test Design and Features on Performance

4:00 to 5:30 pm - Paper Session

Chair:

Fusun Sahin, American Institutes for Research

Participants:

Student and Item Characteristics in Online Reading Comprehension: Polytomous Explanatory IRT Modeling

Hatice Cigdem Bulut, Cukurova University; Serkan Arikan, Bogazici University; Okan Bulut, University of Alberta

This study examines the roles of item and student characteristics on student responses in online reading comprehension (ORC) in an international large-scale assessment. Findings show that ORC performance differences due to gender and speaking the test language at home vary significantly based on item characteristics (e.g., text complexity).

What Matters? Do Device or Examinee Characteristics Drive Score Differences?

Tia Fechter, Office of People Analytics; Daniel Segall, DMDC

This paper explores expanding a large-scale assessment program for administration on a variety of devices including laptops, tablets, and Smartphones. Examinees took two forms of the same test on two randomly assigned devices. Mixed linear effects modeling was used to detect test-level differences for both test scores and response times.

The Effects of Technology-Enhanced Item Formats on Student Performance and Cognition

Burcu Arslan, Educational Testing Service; Blair Lehman, Educational Testing Service

We present the findings and their implications of a within-subject, randomized, controlled experimental study in ELA domain conducted with 535 8th-grade students to investigate the effects of drag-and-drop, inline choice, and grid item formats on student performance and cognition in a sequencing task.

Online Calculator in Large-scale Mathematics Assessments: Usage and Impacts on Item Difficulty

Wei He, NWEA; Patrick Meyer, NWEA

Using data from a large-scale computerized adaptive test, this study examined the use of online calculators by primary and secondary school students in terms of how student background such as gender, achievement level, and special education status affects calculator use and how the use of calculator affects item difficulty.

How Open-Book Assessment Impacts Precision of Person and Item Estimates

Cheng Hua; Stefanie A. Wind, University of Alabama; Stefanie Sebok-Syer, Stanford University

For examining the open-book administration's impact on subject achievement, we used Rasch techniques based on the Rasch measurement theory. The results indicated that outside resources' use impacted the pattern of expected and unexpected responses differently for individual test-takers and individual items. However, the impact is not statistically significant.

Discussant:

Michael Beck, BETA, LLC

Security Issues in Credentialing

4:00 to 5:30 pm - Paper Session

Chair:

Janet Mee, NBME

Participants:

Pandemic is a Portal: A high-stakes medical licensing exam during COVID-19 times

Cecilia Alves, Medical Council of Canada; Maxim Morin, Medical Council of Canada; Nicole Robert, Medical Council of Canada; Becca Carroll, Medical Council of Canada; Allison Burnett, Medical Council of Canada

COVID-19 pandemics has impacted the delivery of many high stakes licensing examinations. In this study, we investigated how examinees' perception of their exam experience in remote proctoring is comparable to test centre proctoring and how examinees' perceived exam experience is related to their overall performance on the exam.

The Impact of Memory on Repeat Item Exposure to Simulation Based Items

Ozge Ersan, University of Minnesota Twin Cities; Matthew Schultz, AICPA

This study examines the impact of item features and memory when examinees' are presented repeat performance assessment items in a high-stakes assessment. Results suggest the benefits of repeat exposure are limited, though some item features produce modest effects. The practical implications to licensing/certification assessments are discussed.

Interaction log Analysis of Proctoring Modalities in high-stakes Medical Licensing Exams

Jinnie Shin; Qi Guo, Medical Council of Canada; Cecilia Alves, Medical Council of Canada; Maxim Morin, Medical Council of Canada

The interaction log comparability between the remote proctoring and the onsite proctoring in the high-stakes medical licensing exam was thoroughly investigated. Results indicated that the proctoring modality may affect the examinee's time-use behaviours in item solving. However, the distinct item solving behaviours showed little to no impact on test performance.

Preknowledge Detection in Multiple-Format Testing

Merve Sarac, University of Wisconsin-Madison; Ting Xu, AICPA

We borrowed information on one format to detect preknowledge on another format within a test. A differential person functioning approach yielded higher power than a regression method. Further investigation revealed that power decreased as the percentage of examinees with preknowledge increased, and the number of contaminated items decreased.

Machine learning algorithms for anomaly detection on Computer-Based Testing *Soo*

Ingrisone, Pearson; James Ingrisone, Pearson VUE

The machine learning (ML) algorithms for anomaly detection on CBT is explored. Hierarchical agglomerative clustering is used for automatically labelling unlabeled data. Random forest ensembles are used to evaluate the accuracy of the clustering. Actual data from a certification exam are used to validate ML classification results.

Discussant:

Michael R Peabody, National Association of Boards of Pharmacy

Remembering "Career" in College and Career Readiness

9:00 to 10:30 am - Organized Discussion

State Departments of Education have generally identified college and career readiness (CCR) measures that were designed for informing college admission decisions. Although academic indicators of English language arts and mathematics are important to career pathways, are these sufficient to characterize the construct of college and career readiness? In this session, participants will discuss topics related to measuring, interpreting, and using career readiness indicators and offer thoughts on how we may rethink current practice.

- Topic 1: Academic indicators as predictors of career readiness.
- Topic 2: Comparable academic indicators to common assessments that measure CCR.
- Topic 3: Applicability of common CCR measures for all students.
- Topic 4: Accountability systems can incorporate career readiness indicators to support equitable interpretation of school or district performance.

The session will include brief overviews from each presenter about defining career in CCR followed by a facilitated discussion by the panel about different perspectives about how CCR has been interpreted and applied in states, and audience Q&A.

Session Organizer & Chair:

Chad W. Buckendahl, ACS Ventures, LLC

Presenters:

Michelle Gough, EdMetric, LLC
Chris Domaleski, Center for Assessment
Alisha Hyslop, Association for Career and Technical Education
David Conely, EdImagine

Diving into NAEP Process Data to Understand Students' Test Taking Behaviors

9:00 to 10:30 am - Coordinated Paper Session

Recently, large scale assessments, including the National Assessment of Educational Progress (NAEP), have transitioned to digitally based assessments (DBAs). Logging timing and behavior data on examinees' interactions with items and delivery interface during the test provides a rich data source, called process data, to examine the relationship between students' testing behaviors and performance. This symposium features three separate studies investigating how process data can be used to identify, classify, and explore students' test taking behaviors, using one block of the 2017 NAEP DBA mathematics grade 4 (N=152,500) administered to a nationally representative sample. The first study examines the non-response patterns (i.e., "omit" and "not reached") in process data and estimate time thresholds of non-response category using machine learning techniques with item-, student-characteristics and process data. The second study uses the time students spend on and between item visits to classify visit behaviors and explore potential motivators for visits, using cluster analysis. The third study analyzes the actions students take within each item visit to identify common action sequences and examine how these differ across items using sequence mining techniques. These studies illustrate how research using process data can contribute to discourses on test assembly, test construction, and test validity issues.

Session Organizer:

Ruhan Circi, American Institutes for Research

Participants:

Revisiting Omit and Not-Reached Scoring Rule using NAEP Process Data

Nixi Wang

Exploring Item Visits in Process Data and Modeling Students' Visit Behaviors

Monica Morell, University of Maryland

Understanding Students' Problem-Solving Processes via Action Sequence Analyses

Manqian Liao

Discussant:

Jonathan Weeks, Educational Testing Service

(Invited Session) Lessons Learned from the Pandemic: How do credentialing programs prepare for the next major crisis/disruption?

9:00 to 10:30 am - Organized Discussion

The year 2020 proved to be an annus horribilis, unleashing an extraordinary combination of forces which placed unprecedented pressures on society – a global pandemic, civil unrest, political tumult, just to name a few. Credentialing bodies were not immune to these pervasive trends and struggled to maintain business continuity and relevance. The unanticipated circumstances and difficulties of 2020 forced certification and licensure bodies to consider options for adjustments and accommodations to credentialing practices, particularly in assessment, that previously were considered risky variations from standard business processes. This session, which features assessment leaders from five national and reputable organizations within the credentialing industry, will explore the unique challenges encountered during 2020 and beyond. The topics explored from the distinct perspectives of these organizations will include: Remote proctoring, transition of examination development workflows to virtual environments, maintaining volunteer and subject matter expert (SME) engagement, candidate accommodations, and business continuity. After exploring these topics, the panelists will turn a cautiously optimistic eye toward forthcoming years and delineate ways in which credentialing practices could be "futureproofed" and made more adaptable to endure through the next major disruption, whatever form that may take.

Session Organizer & Chair:

Timothy Muckle, Board of Pharmacy Specialties

Presenters:

Johnna Gueorguieva
Mary Browne, National Board of Certification and Recertification for Nurse Anesthetists
Sarah Carroll, National Board for Certification in Occupational Therapy
Daniel H. Breidenbach, PSI

Practical Issues in Automated Test Assembly

9:00 to 10:30 am - Coordinated Paper Session

Session Abstract Automated test assembly (ATA) is utilized by many testing organizations to build parallel test forms. In the context of certification and licensure, ATA can also help to ensure that information about examinee performance is maximized around the passing score, which increases classification accuracy. In practice, there are a variety of approaches to implementing ATA and each organization will inevitably have to manage unique challenges for building optimal test forms. This session will have presenters from four different medical certification and licensure boards and current issues they are addressing in automated test assembly. The first paper uses a simulation study to investigate how various proxies for discrimination might be used within a Rasch framework for building exam forms. The second paper examines the realities of combining ATA for assembling an initial test form but often requires the iterative replacement items after review by subject-matter experts and the impact of these practical constraints on form assembly. The third paper focuses on content representation considerations in ATA, specifically focusing on achieving desired, or optimal, levels of content coverage in constructed forms. The fourth and final paper investigates advantages and disadvantages to different open-source platforms for test assembly.

Session Organizer:

Andrew Jones, American Board of Surgery

Participants:

Incorporating discrimination indices into ATA within a Rasch framework

Paulius Satkus, James Madison University; Andrew Jones, American Board of Surgery; Beatriz Ibanez Moreno, American Board of Surgery; Carol L Barry, American Board of Surgery

Automated test assembly in the context of form review and item replacement

J. B. Weir, National Commission on Certification of Physician Assistants; Marcus Walker, National Commission on Certification of Physician Assistants; Joshua Goodman, NCCPA

Content representation considerations in automated test assembly

Robert Thomas Furter, American Board of Pediatrics

Comparison of software platforms for ATA with mixed integer linear programming

Michael R Peabody, National Association of Boards of Pharmacy

Discussant:

Richard Melvin Luecht, UNC Greensboro

Recent Research on Detecting Disengaged Test Taking

9:00 to 10:30 am - Coordinated Paper Session

In recent years there has been a growing interest in disengaged test taking, as new methods for its detection have emerged. In particular, the increased use of computer-based tests has taken advantage of its capability to record item response time. This has been used to identify rapid-guessing behavior, which is a validated indicator of disengaged test taking. Despite its use, however, researchers have continued to seek ways to improve our ability to detect disengagement. Such research has taken multiple paths. One approach has concerned the most effective way to choose the response time thresholds used to identify rapid guessing. Another approach is to investigate ways to detect instances of disengagement that are non-rapid. A third approach is to explore how other forms of process data can be used to identify disengagement. This session's papers, presented by leading researchers in the field, represent each of these approaches. The first three papers investigate the most effective ways to detect rapid guessing behavior using response time information. The fourth paper reports on an innovative method for detecting the presence of non-rapid, partial test-taking engagement. The last paper explores and illustrates how eye-tracking measurement can be used as an alternate methodology for identifying disengagement.

Session Organizer:

Steven Wise, NWEA

Chair:

Dena Pastor, James Madison University

Participants:

Comparing Different Response Time Threshold Setting Methods to Detect Low Effort

James Soland, University of Virginia; Megan Kuhfeld, NWEA; Joseph A. Rios, University of Minnesota

Does the Choice of Response Time Threshold Procedure Matter?

Jiayi Deng; Joseph A. Rios, University of Minnesota

As Testing Contexts Change, How Should Engaged Responding be Identified?

Blair Lehman, Educational Testing Service; Jonathan Steinberg, ETS; Fred Yan, ETS; Jesse R. Sparks, Educational Testing Service; Jung Aa Moon, Educational Testing Service

A Method for Identifying Partial Test-Taking Engagement

Steven Wise, NWEA; Megan Kuhfeld, NWEA

Detecting Test-Taker Disengagement by Means of Eye Tracking: Potentials and Limitations

Marlit Annalena Lindner, IPN Kiel; Burcu Arslan, Educational Testing Service

Communicating results

9:00 to 10:30 am - Paper Session

Chair:

Kristin M. Morrison, Curriculum Associates

Participants:

Improving the Measurement Efficiency of the California School Dashboard

Christopher Cleveland, Harvard University

Using an Item Response Theory Graded Response Model, I find that a more efficient California Dashboard accountability model may rely on ELA and Math performance to identify districts for technical assistance relative to the other existing measures of Chronic Absenteeism, Suspension Rate, English Learner Progress, Graduation Rate, and College/Career Readiness.

Understanding Trends in School Grouping Using Clustering and a Visualization Tool

Steven Tang, eMetric LLC; Zhen Li, eMetric LLC; Zhen Gao, eMetric LLC

This paper investigates K-means, hierarchical, and density-based clustering on real testing data from hundreds of elementary schools from a single state. A "visual clustering" approach is proposed to allow stakeholders to engage with the clustering in real-time. Results from real-data analysis will be presented.

Evaluating Usability and Utility of a Teacher Dashboard to Support Instructional Decision-Making

Robert Dolan, Diverse Learners Consulting; Kim Ducharme, CAST; Samantha Gilbert, SM Education; Allison Posey, CAST

A small-scale study was conducted to evaluate a prototype teacher dashboard designed to provide teachers with science assessment results contextualized within dynamic learning map (DLM) models to support instructional decision-making. Findings indicate high degrees of both usability and utility of the dashboard and hence promise for supporting effective formative assessment.

State Assessment Score Reporting Practices for Limited English Proficient Parents

Samuel Dale Ihlenfeldt, University of Minnesota; Joseph A. Rios, University of Minnesota

This study investigated nationwide whether score reporting for state accountability and ELP assessments (e.g., WIDA) was accessible to parents of English learners (i.e., translated and following best score reporting practices). Results indicate differences between assessment types, as well as key trends across both. Implications for practice are discussed.

Conveying Uncertainty in Score Reports

Zhaopeng Ding; Mark Hansen, UCLA

Interpretations and uses of test results may depend in part on how stakeholders understand uncertainty in scores and classifications. The present study examines whether different visual representations of uncertainty affect educators' interpretation of test results and, if so, whether these effects are moderated by educators' characteristics.

Discussant:

Bryan R. Drost, Rocky River Schools

The Impact of COVID-19 on Educational Measurement, Part 2: Admissions and Certification

9:00 to 10:30 am - Coordinated Paper Session

In the summer of 2020, Educational Measurement: Issues and Practice editor Deborah Harris invited selected testing and measurement experts to assess the impact of the COVID-19 pandemic on the field for a special issue of the journal. In two separate symposia, contributing authors to this special issue will present their evaluation of the impact of COVID-19 on the profession, including their best ideas about how the educational measurement community can and should respond to these unprecedented challenges. A first symposium includes papers related to K-12 assessment. This second symposium includes papers related to admissions and certification testing. Presentations will be short, emphasize cross-cutting themes, and leave plentiful time for audience questions.

Session Organizer:

Andrew Ho, Harvard Graduate School of Education

Chair:

Ye Tong, Pearson

Participants:

Never Let a Crisis Go to Waste: Large-Scale Assessment and the Response to COVID-19

Wayne J. Camara, LSAC

Standardized Testing in College Admissions: Observations and Reflections

Li Cai, UCLA

Impacts of COVID-19 on the Law School Admission Test

Lily Knezevich, Law School Admission Council; Josiah Evans, Law School Admission Council

Internet-Based Proctored Assessment: Security and Fairness Issues

Thomas E. Langenfeld, TEL Measurement

When Examinees Cannot Test: The Pandemic's Assault on Certification and Licensure

Michael Jodoin, National Board of Medical Examiners; Jonathan Rubright, National Board of Medical Examiners

Your Guess is as Good as Ours

Andrew Wiley, ACS Ventures, LLC; Chad W. Buckendahl, ACS Ventures, LLC

Discussant:

Andrew Ho, Harvard Graduate School of Education

Friday Coffee Chat Sessions

10:35 to 11:00 am

Join your NCME colleagues for a unique opportunity to share ideas, questions, and thoughts about current topics in our field. Or enjoy a brief time for relaxation!

1. **Coffee Chat: Future of K-12 Assessment**
   Hosted by Randy Bennett, ETS

2. **Coffee Chat: Fighting for Fairness - Where do we go next?**
   Hosted by Thanos Patelis, Fordham University, Teachers College, University of Kansas

3. **Coffee Chat: Opportunities with Response and Process Data**
   Hosted by Arslan Burcu, Educational Testing Service

4. **Coffee Chat: Going for Broke: Truly Equitable Assessments Require Disrupting the White Supremist Status Quo**
   Hosted by Jennifer Randall, University of Massachusetts, and Kristen Huff, Curriculum Associates

5. **Coffee Chat: Federal Peer Review: Opportunities and Challenges for Educational Measurement**
   Hosted by Brian Gong, Center for Assessment

6. **Coffee Chat: Chat with NCME's newest Past President**
   Hosted by Ye Tong, Pearson

7. **Coffee Chat: Chill with a little Chat**
   Give your mind a little rest in this 25-minute break. We'll say hello and settle in for the first 5 mins, have a guided meditation for 15 mins, and use the last 5 mins for a little chat before we transition to the next session.
   Hosted by: Rosemary Reshetar, National Conference of Bar Examiners

Developing a Longitudinal Assessment: Using Innovations and Research to Address Measurement Issues

11:15 to 12:45 pm - Coordinated Paper Session

The landscape of education is continuously evolving; in turn the field of educational measurement must evolve with it. Measurement practitioners must learn to adapt their traditional measurement processes to more appropriately align with the future of educational assessment, measurement, and the needs of the stakeholders. One example of how educational assessment has progressed due to stakeholder needs is the transformation from traditional point-in-time assessments to longitudinal assessments in medical certification. The shift to this new assessment structure is to allow examinees flexibility when taking their assessments, while maintaining rigorous measurement principles. This session is beneficial for testing organizations planning to move from a traditional point-in-time assessment to a continuous assessment structure, or to any testing organization interested in longitudinal assessment. Several measurement challenges faced with this type of assessment change, and the research and innovations being conducted to address them, are covered. The first paper in this session contrasts the traditional point-in-time assessment with a more longitudinal assessment design, which sets the stage for the remaining four papers: (1) pretesting and equating, (2) form assembly, (3) scoring, and (4) ensuring test security.

Session Organizer:

Pamela Kaliski, ABIM

Participants:

Setting the Stage: An Overview of ABIM's Longitudinal Assessment Program

Pamela Kaliski, ABIM; Whitney Coggeshall, American Board of Internal Medicine; Jerome Clauser, American Board of Internal Medicine

Considerations for Pretesting and Equating Within a Continuous Assessment Framework

Kelly Rewley, American Board of Internal Medicine; Jerome Clauser, American Board of Internal Medicine; Kelli Samonte, American Board of Internal Medicine; Deirdre Derrick, American Board of Internal Medicine

Balancing Item Attributes in a Longitudinal Assessment

Whitney Coggeshall, American Board of Internal Medicine; Jerome Clauser, American Board of Internal Medicine

Determining a Scoring Approach for the ABIM Longitudinal Assessment Program

Whitney Coggeshall, American Board of Internal Medicine; Kelli Samonte, American Board of Internal Medicine; Pamela Kaliski, ABIM

Security and the ABIM Longitudinal Assessment Program

Derek Sauder, American Board of Internal Medicine; Jin Zhang, American Board of Internal Medicine

Discussant:

Michael Kane, ETS

Electronic Board Session #3

11:15 to 12:45 pm

Participants:

A Machine Learning Method for Classify Student's Learning Status

Zhemin Zhu, Beihua University; Hua-Hua Chang, Purdue University

It is challenging to acquire students learning status when the response data is contaminated by aberrant patterns. This research developed a machine-learning method for Cognitive Diagnostic Models to achieve the goal. A pilot simulation showed that the new method is more effective than some traditional ones, especially for short-length tests.

An Application of Topic Modeling for Investigating Mathematics Teachers' Reasoning

Minju Hong, University of Georgia; Yasemin Copur-Gencturk, University of Southern California; Hye-Jeong Choi, University of Georgia; Allan Cohen, University of Georgia

This study extends the structural topic model to include both covariates and outcome variables. An illustrative example is presented showing how the model can be used to detect the latent thematic structure in test answers and the relationship of that structure to the scores on the test.

An Investigation in UAMIRT on Testlet-Based Tests Equating under a NEAT Design

Qianqian Pan, University of Hong Kong; Hongyu Diao, University of Massachusetts Amherst

This study investigates the performance of unidimensional approximation of MIRT model on testlets equating under current calibration and separate calibration designs.

Application of Multilevel Modeling in Large-scale Assessments: A Systematic Review

Olasunkanmi Kehinde, Washington State University; Shenghai Dai, Washington State University; Brian French, Washington State University

Educational researchers and practitioners are often hesitant about using multilevel modeling (MLM) to analyze large-scale assessment (LSA) data due to various reasons such as the methodological challenges in both frameworks. We conduct a systematic review of the education literature to inform MLM applications in LSAs.

Assessing Differential Item Functioning Flagging Rules Using a Sample of Examinees

Tzu-Chun Kuo, Cambium Assessment, Inc.; Tao Jiang, Cambium Assessment, Inc; MinJeong Shin, American Institutes for Research

Four sampling methods were compared in analyzing DIF using the Generalized Mentel-Haenszel procedure. Preliminary results indicated that stratified sampling by group, item score, and number of contingency subtables is preferred, and that larger sample size is needed if the two groups have larger DIF.

Detecting Differential Item Functioning Using the Bayeisan Factor Method

Nan Wang

Differential item functioning(DIF)analysis is an important practice in verifying the validity of an exam in the field of education. In the present study,I propose to use Bayesian factor method within the Mantel-Haenszel framework to detect the DIF among different groups of students.

Effect sizes for estimating differential item functioning influence at the test level

Holmes Finch, Ball State University; Brian French, Washington State University

The understanding of the cascading influence of differential item function on test scores used to make decisions is critical to fair and equal outcomes for all individuals. Several proposed effect size measures that capture this DIF influence are compared through simulation and applied to a real dataset. Implications are discussed.

Evaluating Computer-based Test Accommodation for Students with Disabilities using DIF models

Haeju Lee; Kyung Yong Kim, University of North Carolina at Greensboro; Hongwook Suh, Nebraska Department of Education; Stephen Sireci, University of Massachusetts Amherst

In this study, we investigate whether items function differently across groups of students that took a computerized-adaptive math test with and without Text-to-Speech accommodation. Logistic regression and Item Response Theory based likelihood ratio test are used to detect differential item functioning.

Evaluating Stability and Psychometric Properties of Vertical Scale Scores

Yi-Fang Wu, ACT

Methodologies of vertical scaling have been extensively discussed but approaches to evaluating vertical scales in existence requires more investigation. This study evaluates stability and psychometric properties of scores on two existing vertical scales when data collection designs are less optimal than the original designs for establishment of the scales.

Exam Item Reduction: Classification Accuracy Study

Igor Himelfarb; Guoliang Fang, Colorado State University Global; Nai-En Tang, National Board of Chiropractic Examiners

This study was conducted to investigate classification accuracy between the full and reduce forms of the Chiropractic Clinical Sciences Exam (Part II) using methodologies based on classical test theory (CTT), item response theory (IRT) and Bayes' Theorem. Results showed high accuracy consistent across all three methods.

Exploring Invariance of Classroom Practices Across Countries: Chile, Colombia and Mexico

Mariana Barragan Torres, UCLA

Using large-scale international data has allowed exploring outcomes across countries. However, comparisons rely on the assumption of invariance. To test this assumption, I explore the factorial structure of six domains of classroom practices in the Talis Video Study for two measures: observation and student ratings in Chile, Colombia and Mexico.

Impacts of COVID-19 Pandemic on Noncognitive Variables of Grade Five Students

Jun Li, University of Minnesota Twin Cities; Qian Zhao, University of Minnesota Twin Cities; Julio Caesar, Bloomington Public Schools

Schools in the United States were widely closed due to COVID-19 pandemic. With the survey data of 960 grade 5 students, we used item response theory unidimensional framework, t-test, and Chi-square to estimate differences in noncognitive variables between 2019 and 2020, and found some significant changes after the school closure.

Measuring in-platform learning in online learning systems that support formative assessment

Jinnie Choi, Savvas Learning Company; Yun Jin Rho, Maguire Associates; Emily Lai, Pearson

How do we measure 'learning' that happens while learners engage in formative assessment in online learning systems (where learners can use hints and multiple attempts to get correct answers)? We propose in-platform measures of mastery that can reveal practical insights about improving teaching and learning with formative assessment.

Merging Multiple Sources of Evidence to Improve Score Results

Eunhee Keum, UCLA CRESST; Mark Hansen, UCLA; Preston Botter, UCLA CRESST

It is widely agreed that high-stakes decisions should not be based on a single test result, yet such practice abounds. One potential barrier to using multiple results is uncertainty in how evidence should be combined. We explore whether a latent regression item response theory model is suitable for this purpose.

Modeling Change in PA General Medical Knowledge Over Time

Jennifer L. Lewis, University of Massachusetts Amherst

The current study explores the performance of physician assistants across six quarters of an alternative recertification exam. This study aims to gather more information about changes in performance over time within a test-enhanced learning context.

Psychometric Analyses of the TIMSS Exam Using Generalizability Theory

Jaime Malatesta, Graduate Management Admission Council; Tong Wu; Stella Kim, University of North Carolina at Charlotte; Won-Chan Lee, University of Iowa

This study aims to 1) examine the psychometric properties of the TIMSS assessment using a Generalizability Theory framework, 2) compare results of several univariate and multivariate designs, and 3) discuss differences between the methodology used in the current study and that used operationally for the TIMSS assessments.

Quality Assurance in Through-Course Assessment: An Evaluation Plan

Kun Su, UNC Greensboro; Shonai Someshwar, UNC Greensboro

This study intends to develop and showcase the application of an evaluation plan for the through-course assessments. It aims to illustrate the underlying assumptions of the system and the statistical procedures to be implemented to help other test developers using similar designs in evaluating their own systems.

The Impact of the Within-family Variation of Item Parameters on Scoring Precision

Chen Tian; Jaehwa Choi, George Washington University

Within an item family, automatic item generation creates instances sharing similar psychometric characteristics. This study explores how the within-family variation of item parameters affect scoring when identical item sibling parameters were assumed. Results show that the scoring precision was just slightly undermined in both CAT and linear tests.

The Use of Posterior Probability for Score Differencing

Sandip Sinharay, Educational Testing Service; Matthew Johnson, ETS

A Bayesian approach was suggested for score differencing. The approach involves the computation of the posterior probability of a better performance on one subtest compared to another. The new approach leads to fewer false alarms and more true positives compared to existing approaches in a simulation study.

Time Use on the Shortened MCAT® Exam Under COVID-19

Ying Jin, Association of American Medical Colleges; Marc Kroopnick, Association of American Medical Colleges; Monica Morell, University of Maryland

The MCAT exam was shortened in response to COVID-19. This study examined how examinees used their time on the shortened exam relative to its full-length version. The findings suggest that examinees spent similar amount of time on content common across versions and tended to engage with the two versions similarly.

Using Machine Learning to Administer Salt Items in Computerized Adaptive Testing

Zhongmin Cui, CFA Institute; Chunyan Liu, National Board of Medical Examiners; Yong He, ACT

Computerized Adaptive Testing with Salt (CATS) implements CAT with unrestricted item review and answer changes while being robust to cheating strategies. A successful implementation of CATS depends on effective administration of salt items to the right test takers. Machine learning was found to be helpful on improving the effectiveness.

Using the Teamwork Expectations and Attitudes Measure (TEAM) to Assess Student Perceptions of Working in Teams

Brandon Johnathan Justus; Shayna Rusticus, Kwantlen Polytechnic University; Brittney Stobbe, Kwantlen Polytechnic University; Jonathan Lau, Kwantlen Polytechnic University

We developed a measure of attitudes towards working in a team called the Teamwork Expectations and Attitudes Measure (TEAM). Two pilot studies refined the TEAM into a 14-item unidimensional scale. A third validation study confirmed the unidimensional structure and provided evdence of convergent, discriminant, and criterion validity.

Validation and adaption of the PATHS scale by Chinese preschoolers

Menglong Cong, University of Denver

The Promoting Alternative THinking Strategies (PATHS) program is a Social-Emotional Learning program. The validation of the PATHS evaluation tool by Chinese preschoolers is under-researched. Results indicated that the three-factor correlational model has adequate model fits by Chinese preschoolers and Scalar invariance achieved by gender. Scale adaption suggestion was provided.

Virtual vs. In-person Standard-Setting for High-Stakes Medical Performance Assessments: Comparing Validity Evidence

Andrea Julie Gotzmann, Medical Council of Canada; Fang Tian, Medical Council of Canada; Sirius Qin, Medical Council of Canada; Yousef Mousavi, Medical Council of Canada; Lucy Zhang, Medical Council of Canada

Evaluating post exercise surveys (procedural evidence) and Generalizability analyses of ratings (internal evidence) from a virtual and recent in-person standard-setting exercises on two different high-stakes clinical performance assessments. The two performance assessments will be using contrasting groups and borderline group methods.

Identifying Rushing in CAT and Investigating the Effects on Differentiated Instruction

11:15 to 12:45 pm - Coordinated Paper Session

To attain test scores that validly indicate what a student knows and can do, students must exhibit motivated and effortful behavior throughout the testing event. Therefore, it is pivotal to accurately identify non-effortful behavior. Computer based testing has made possible the use of response time-based measures for separating effortful from non-effortful behavior. This symposium conceptualizes non-effortful behavior as rapid-guessing, e.g. rushing, at the item and test level. All of the papers focus on identifying rapid-guessing on the i-Ready Diagnostic, an interim CAT taken three times a year by students in kindergarten through twelfth grade in reading and mathematics. The first paper in this symposium introduces frameworks for separating rapid-guessing behavior from effortful behavior in an operational setting. The second paper builds upon the first and outlines improvements based on monitoring initiatives. The third paper investigates the practical implications of using rushed assessments to route students through online instruction and provides some initial validity evidence for the current rush flags implemented in the i-Ready Diagnostic.

Session Organizer:

Alexandra Lay

Participants:

Developing Item- and Test-Level Rush Flags for the i-Ready Diagnostic Assessment

Elizabeth Adele Patton; Logan Rome, Curriculum Associates

Monitoring and Improvement of the i-Ready Diagnostic Rush Flags

Elizabeth Adele Patton; Logan Rome, Curriculum Associates; Alexandra Lay

Evaluating Practical Implications of Using Rushed i-Ready Diagnostic Scores for Instructional Purposes

Alexandra Lay; Elizabeth Adele Patton; Logan Rome, Curriculum Associates

Designing and Evaluating Innovative Assessment Systems: Combining Research and Practice

11:15 to 12:45 pm - Coordinated Paper Session

State assessment programs are rarely evaluated using program evaluation approaches. An alternative is to more closely connect assessment design to evaluation mechanisms and, in doing so, draw on the traditions of educational measurement and program evaluation. In bridging these traditions, this coordinated session features research and evaluation related to the practical measurement challenges and technical issues in the context of implementing federally-approved statewide innovative assessment systems. These systems provide one ideal context to illustrate tighter connections between measurement and evaluation because these innovative systems are often designed with action mechanisms directly aimed at the classroom. The session will begin with presenting an evaluative framework for state-level programs that are intended to inform instruction in addition to supporting systems of school identification required under federal law. The subsequent presentations on New Hampshire and Louisiana will connect to this evaluative framework and serve as case study examples of research-practice agendas around outcomes research, measurement considerations, and technical issues. The presentation on New Hampshire's innovative system highlights findings from an outcome evaluation after five years of implementation. The presentation on Louisiana's innovative system explores rater error and its impact on summative determinations.

Session Organizer:

Carla M. Evans

Participants:

Balancing Rigor and Relevance: A Framework for States to Evaluate and Improve Innovative Assessment Systems under IADA

Chris Brandt, Center for Assessment; Nathan Dadey, Center for Assessment; Carla M. Evans

NH Performance Assessment of Competency Education Student Outcome Evaluation after Five Years

Alexandra Stone, University of Connecticut; Carla M. Evans

Understanding the Impact of Rater Inaccuracies on Test Score Scales

Tong Wu; Michelle Boyer, Center for Assessment

Discussant:

Matthew N. Gaertner, WestEd

The Value of and Values in Educational Assessment

11:15 to 12:45 pm - Coordinated Paper Session

Values underlie all aspects of the establishment, development, and evaluation of educational testing programs. To address current criticisms of educational tests; such as narrowing the curriculum, widening achievement gaps, focusing on unimportant skills, and creating anxiety in children; we will discuss the values inherent in current educational testing programs, and how we can reexamine these values to promote more equitable outcomes for students. The symposium will consist of two parts. The first part will comprise three 10-minute presentations on current values in educational assessment policy, educational test development, and educational test evaluation. The second part will be a blue-ribbon panel discussion of those topics focused on improving educational equity. Specifically, the panel will respond to the question "How can we re-center our current values in educational assessment to empower traditionally marginalized groups in the educational assessment process?" Dialogue among the panelists, presenters, and the audience will be facilitated. A goal of the symposium is to discover ways in which educational testing can be re-conceptualized to support the learning and progress of all students.

Session Organizer:

Stephen Sireci, University of Massachusetts Amherst

Chair:

Cindy M Walker, Research Analytics Consulting LLC

Participants:

What are our current values in Educational Assessment Policy?

Sujie Shin, California Collaborative for Educational Excellence for Carl Cohn

What are our current values in educational test development?

Maria Elena Oliveri, Buros Center for Testing-UNL

What are our current values in educational test evaluation?

Stephen Sireci, University of Massachusetts Amherst

Panel Discussion

Stafford Hood, University of Illinois at Urbana-Champaign

Suzanne Lane, University of Pittsburgh

Darius Prier, Duquesne University

Jennifer Randall, University of Massachusetts

Amy Stuart Wells, Teachers College

PISA and TIMSS Topics

11:15 to 12:45 pm - Paper Session

Chair:

Anthony Albano, University of California, Davis

Participants:

Performance of Multi-group DIF Methods in Assessing Cross-Country Comparability of TIMSS's Math Scores

Dandan Chen, University of Illinois at Urbana-Champaign; Jinming Zhang, University of Illinois at Urbana-Champaign

The vast majority of the literature on DIF methods is limited to two groups. Two recent multi-group DIF detection methods have been developed to detect both uniform and nonuniform DIF among more than two groups. These two methods are the improved Wald test and the generalized logistic regression procedure. This study assessed the commonalities and differences between two sets of empirical results from these two methods with the latest TIMSS math score data from six countries. The primary conclusion was that the improved Wald test might be relatively more established than the generalized logistic regression procedure for multi-group DIF analysis.

Investigation of Item Bias on the PISA 2009 Reading: Chinese and English Versions

Sok-Han Lau, University of Hawaii at Manoa; Seongah Im

The purpose of this study was to investigate possible item bias exist in the PISA 2009 reading assessment of Macao. Differential item functioning (DIF) using the Mantel-Haenszel (MH) method and item response theory (IRT) statistics using the improved Wald Test were conducted. Both statistical methods identified the same DIF items.

Is Item Disengagement Different Across Years? Comparisons between PISA 2018 and 2015

Huan KUANG, University of Florida; Fusun Sahin, American Institutes for Research

Disengaged responses can bias item estimations. Whether this bias is similar across administrations is unknown. We detected disengaged responses from computer-based mathematics common items in PISA 2015 and 2018. Up to 3% of responses were detected as disengaged. Effects of removing disengaged responses on item difficulty were reported.

Flexible Modeling of Item Responses and Response Time Using Splines

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

We analyze the 2015 PISA mathematics data and conclude that routinely used parametric factor models for response accuracy and time yield unsatisfactory fit and misleading factor scores. We propose a semiparametric approach with spline-based functional estimators that can model flexibly the distribution of latent and observed variables.

Automatic Normalization for Advancing Response Coding Consistency and Efficiency in PISA

Fabian Zehner, DIPF | Leibniz Institute for Research and Information in Education, Centre f. Int. Student Assessm.; Hyo Jeong Shin, Educational Testing Service; Emily Kerzabi, Educational Testing Service; Nico Andersen, DIPF | Leibniz Institute for Research and Information in Education; Frank Goldhammer, DIPF | Leibniz Institute for Research and Information in Education, Centre f. Int. Student Assessm; Kentaro Yamamoto, Independent Researcher

Many items in PISA require scoring text responses by humans. For scoring automatically, we propose normalization techniques that improve their grouping. The study focuses on scoring consistency and efficiency. Our investigation of 2.5 million text responses and 14 country-by-language groups demonstrates improvements for both with a minor loss in accuracy.

Discussant:

Jon S. Twing, Pearson

Adaptive Testing Topics

11:15 to 12:45 pm - Paper Session

Chair:

Anne Traynor, Purdue University

Participants:

Strategies for Implementing CD-CAT in High-Dimensional Testing Situations

Yan Sun, Rutgers University; Jimmy de la Torre, University of Hong Kong

CD-CAT has been developed to administer diagnostic tests more efficiently; however, when the number of attributes is large (i.e., test is high-dimensional), implementing CD-CAT becomes infeasible. To address the high-dimensionality issue, a strategy that involves item calibration, modified item selection and shrinking the prior distributions is proposed.

A Time Constrained CAT Design to Support Online Testing

Tong Wu; Hua-Hua Chang, Purdue University

With the COVID-19 pandemic, many schools have to cancel classes and move to online instructions which may cause tremendous challenges for teaching and testing. The objective of research is to propose a Response Time (RT) constrained online Computerized Adaptive Testing (CAT) for students to take high-stake exams at home possible.

A Comparison of Final Scoring Methods under the MST Framework

Hacer Karamese; Won-Chan Lee, University of Iowa

The purpose of this simulation study is to investigate the performance of final scoring methods under the 1-3 and 1-3-3 MST designs. Simulations are performed to compare scoring methods and panel designs.

Impact of Restricting Grade Level Field-Testing in a Vertically Scaled CAT

Nathan Wall, eMetric; Katherine Nolan, Curriculum Associates; Aimee Boyd, Curriculum Associates

Three field-test design conditions were studied to evaluate the impact upon item parameter estimates while varying grade level ranges for operational items and grade level ranges for students. Implications of this research serve to inform future field-test calibration designs for adding items to a CAT's vertically scaled item pool.

Using CTT Classification Consistency Estimates in a CAT Context

Ramsey Cardwell

Classification consistency (CC) communicates score stability but requires conditional standard errors to estimate in IRT–CAT contexts. A CAT simulation study and real-data example comparing CTT–based CC estimates found all methods consistently underestimated true CC, but some exhibited small bias. CTT-based methods can thus provide convenient and interpretable conservative CC estimates.

Discussant:

G. Gage Kingsbury

(Invited Session) Looking ahead – Bridging future research and practice in credentialing

1:00 to 2:00 pm - Organized Discussion

In this invited session we challenge panelists and the audience to critically evaluate the current research landscape and consider how research can better inform practice in certification and licensure. Practice should be grounded in research; however, the real-world conditions for practitioners are often less than ideal and can limit the applicability of research. Barriers can include but are not limited to organizational constraints, resources, and a lack of expertise within specific domains of research. This leads to situations in which use cases are limited and practitioners struggle to operationalize research, which in turn causes well-defined research to be under-utilized. This disconnect adversely impacts both researchers and practitioners, leaving neither well-served. In this session, sponsored by the Certification & Licensure SIGIMIE, panelists will discuss the disconnect between research and practice with the goal of identifying specific actions that can be taken to increase alignment within the context of credentialing testing.

Session Organizer:

Andrew Jones, American Board of Surgery

Chair:

Michael R Peabody, National Association of Boards of Pharmacy

Presenters:

Deborah Harris, University of Iowa
Andre Rupp, Mindful Measurement
Rich Feinberg, National Board of Medical Examiners
Ada Woo, Ascend Learning

Focus on CDM and DCM

1:00 to 2:00 pm - Research Blitz Session

Chair:

Kelley Wheeler, ACS Ventures, LLC

Participants:

Evidence-Based Feedback in Higher Education through Constructive Alignment and Cognitive Diagnostic Modeling

Stefan Behrendt, University of Stuttgart

Constructive Alignment provides a theoretical model for competence-oriented teaching in higher education. Defining the learning objectives as skills and using the exam's tasks, one can apply Cognitive Diagnostic Modeling methods. This talk investigates practical implications of this approach for teaching and learning, focusing on feedback processes in mechanical engineering courses.

Relative Robustness of CDMs and (M)IRT in Measuring Change in Latent Skills

Qi Huang, University of Wisconsin-Madison; Daniel Bolt, University of Wisconsin-Madison

We examine the relative robustness of longitudinal applications of CDMs and (M)IRT in measuring growth in latent skills, focusing on the performance of each method in data generated under conditions from the alternative model. The results suggest that (M)IRT shows greater robustness.

Using Cognitive Diagnostic Analysis to Construct Learning Path of Data-Analysis-Knowledge for Pre-Service Teachers

xiaopeng wu; Xiuxiu Tang; Hua-Hua Chang, Purdue University

This study develops a method for assessing pre-service teachers' data analysis knowledge from a literacy survey. Among an array of cognitive diagnosis models, the GDM model outperformed regarding model-data-fit and was selected to report score attribute, mastery probability, and learning path based on a real dataset.

A Study for General Diagnostic Classification Model Under Conditions of Extreme Base Rates

Meina Bian, University of Georgia; Laine Bradshaw, University of Georgia

This study investigates the general diagnostic classification models (GDCMs) under conditions where mastery rates are particularly high or low. Results inform the extent to which GDCMs are accurate under increasingly complex educational measurement scenarios.

Context Matters: A Comparison of Empirical CDM Analyses Involving Two Different Q-Matrices

Qianru Liang, University of Hong Kong; Kevin Carl Santos, University of the Philippines; Hartono Tjoe, Penn State University; Jimmy de la Torre, University of Hong Kong

This study investigates how Q-matrices developed using two different curricula (i.e., US and Hong Kong) affect model-data fit and inferences about examinees' mastery profiles. Analyses of proportional reasoning test data collected in Hong Kong show that, although the two Q-matrices provide acceptable model-data fits, their examinee classifications are dramatically different.

Comparison of Matching Criteria of DIF Detection Methods in the CDM Framework

Gamze Kartal, University of Illinois at Urbana-Champaign; Jinming Zhang, University of Illinois at Urbana-Champaign

Cognitive diagnostic models have gained growing attention, but differential item functioning is still an issue in the CDM framework. This study compares the performances of two DIF detection methods, MH, LR, under two different matching criteria, total score and attribute pattern to check the validity and fairness of an assessment.

Explanatory Modeling of Language DIF Using the IRT-C and E-CDM Models

Kevin Krost, Fralin Biomedical Research Institute

Differential item functioning (DIF) was evaluated between English- and Spanish-speaking students on released science items from 2011 TIMSS using IRT and CDMs. For items exhibiting DIF, covariates were modeled to explain DIF. Last, item content features were evaluated to explain any remaining DIF after modeling the covariates.

Dealing with Nonignorable Missingness Assuming Variable Speed in Cognitive Diagnostic Modeling

Yi Yang, Columbia University; Yi Chen, Teachers College, Columbia University; Young-Sun Lee, Teachers College, Columbia University

Fox and Marianti (2016) modeled variable speed utilizing response accuracy (RA) and response time (RT) simultaneously while most joint models assume constant speed. This paper will explore its extension in Cognitive Diagnostic Modeling (CDM), and investigate its application in modeling missing data process using a fully Bayesian approach.

Item Response Theory Applications

1:00 to 2:00 pm - Research Blitz Session

Chair:

Eunbee Kim, Georgia Institute of Technology

Participants:

Constructing a Common Grade Point Average Scale Using Advanced Placement Exams

Weiwei Cui, College Board; Michael E. Walker, Educational Testing Service

In this study, we construct a common scale for GPAs within and across schools in the United States using Advanced Placement (AP) exam scores as anchors to eliminate course and school effects. Item response theory based modeling is used as a scaling method to adjust AP scores across AP exams.

Simulation Study: Evaluating Rater Category Ordering with the JML-Rasch-MFRM Model in Facets

Chunling Chunling Niu, University of Kentucky; Kelly Bradley, University of Kentucky; Shannon Sampson, University of Kentucky; Rui Jin, University of Kentucky; Yuyan Xia, University of Kentucky; Nan Li, University of Kentucky; Lijun Shen, University of Kentucky; Rongxiu Wu, University of Kentucky; Jing Zhang, University of Kentucky

The Rasch-MFRM approach has been employed to detect and measure various rater effects, but each rater is assumed to rate based on the ordered rating scale categories as intended. Thus, this simulation study aims to investigate the diagnostic sensitivity of Rasch-MFRM towards the rating scale category disordering for individual raters.

Validating Scoring Rubrics of Scientific Inquiry Tasks Using Item-Response Theory

Tao Gong, Educational Testing Service; Xiang Liu, Educational Testing Service; Adrienne Sgammato, ETS

We propose a post-hoc method to validate predefined scoring rubrics of scientific inquiry tests using item-response theory and response strategies. Based on two tasks in the National Assessment of Educational Progress, we implement this method and show that it helps validate and modify scoring rubrics to classify students' capacities.

Evaluation of the MH-RM algorithm for the crossed random effects model

Jia Hao, University of Minnesota Twin Cities; Seungwon Chung

Estimation of generalized linear mixed effects model with crossed random effects is computationally challenging. This study proposes MH-RM as a qualified alternative to MCMC for this class of models, especially when sample size is small. The sensitivity of various priors on estimation accuracy and stability is also examined.

A construct modeling approach to the performance assessment of teaching practice

Amy Dray, Spencer Foundation; Diah Wihardini, Bina Nusantara University; Mark Wilson, University of California, Berkeley; Pamela Moss, University of Michigan

This paper provides a description of a performance assessment of two teaching practices—leading a classroom discussion and assessing students between and within lessons. Construct modeling is used to elucidate learning progressions in teacher practice. The system of assessment is described and analyses of teacher enactments of practice are discussed.

The Effect of Option Similarity on the Item Difficulty of a Reading Comprehension Test: An Application of Word Embedding Techniques

Tahereh Firoozi; Jinnie Shin; Okan Bulut, University of Alberta; Mark Gierl, University of Alberta

This study investigated the effect of semantic similarity of the options on the difficulty level of a reading comprehension test items. Semantic similarity of the options was captured using natural language processing techniques and its contribution to item difficulty was estimated using different measurement models. Results indicated that semantic similarity of the options explains more than half of the variance in the item difficulty level.

Examining Examiner Bias using Many-Facet Rasch Model

Nai-En Tang; Chia-Lin Tsai; Igor Himelfarb; Andrew Gow, National Board of Chiropractic Examiners

Examiner bias (Fuchs & Fuchs, 1986) may influence examinee's scores in a clinical performance-based exam (Guraya et al., 2010). Many-Facet Rasch Model (Linacre, 1991) was applied to a 10-station chiropractic technique exam rated by 80 examiners for 437 examinees. The results suggested some variability among examiners. Fair scores were calculated.

Measurement of Transacademic Skills

1:00 to 2:00 pm - Research Blitz Session

Chair:

Chris Brandt, Center for Assessment

Participants:

The Developmental Curve of Students' Social-Emotional Competencies and Practical Implications

Yang Caroline Wang, Education Analytics; Robert H Meyer, Education Analytics

We apply multilevel growth curve models to longitudinal self-report survey data since 2014-15 to examine student SEL developmental patterns and differentiations among student subgroups in grades 3-12. We also discuss practical implications for measurement professionals and educational practitioners as we absorb the impact of COVID from a whole child perspective.

Measurement Invariance Across Student Identity Intersectionalties on SEL Measures, an EIRM Approach

Michael Dosedel, University of Minnesota

Students' social-emotional learning continue gaining prevalence in educational discussions and reform efforts. In this study, explanatory item response modeling is used to evaluate item-level measurement invariance in polytomous items based on the Developmental Asset Profile, exploring the benefits and drawbacks of utilizing Explanatory Partial Credit Models, verses simpler methods.

Effects of Knowledge of Results Feedback Modalities in Testing: A Large-Scale Experiment

Livia Kuklick, IPN Kiel; Marlit Lindner, IPN Kiel

We experimentally varied the modality of knowledge of results feedback during a low-stakes test to investigate feedback effects on learning, test-taking motivation and achievement emotions. Feedback had a small positive effect on posttest performance, whereas effects of feedback on motivational and emotional measures were a function of students' test performance.

Knowledge Acquisition in Higher Education Economics and its Relation to Students` Confidence

Jasmin Schlax, Johannes Gutenberg University; Olga Zlatkin-Troitschanskaia, Johannes Gutenberg University; Marie-Theres Nagel, Johannes Gutenberg University

While the use of multiple-choice tests in higher education is increasing, this assessment type provides only limited indications of students' response processes. Self-reports on confidence in test responses provide insight into these processes, especially guessing behavior. This study investigates the relationship over time between self-confidence and MC knowledge test scores.

A Text-Mining Approach to Measuring Creativity

Denis Dumas, University of Denver; Peter Organisciak, University of Denver; Michael Doherty, Actor's Equity Association

The reliable and valid measurement of creative thinking has been an elusive goal of psychometricians for decades. This paper details the application of semantic-network algorithms to the quantification of human creativity and compares the psychometric properties of scores derived from various text-mining models to those from human raters.

Validating a Liberal Arts Inclination Scale: Applications in classical and modern test theory

Gabe Avakian Orona, University of California, Irvine; Richard Arum, University of California, Irvine; Jacque Eccles, University of California, Irvine; Andrew Maul, University of California, Santa Barbara

Higher education continues to seek ways of expanding assessments beyond cognitive and labor market outcomes to evaluate the effects of college. This study introduces a novel liberal arts inclination instrument and performs preliminary classical, modern, and predictive validity techniques across a range of affective, behavioral, and cognitive outcomes.

Development of personality instrument under the Five-Factor Model for Brazilians

Juliana Cerentini Pacico, University of Iowa; Hacer Karamese, University of Iowa

There is not an accurate instrument to measure personality in Brazil according to FFM- Five-Factor Model. This presentation aims to develop an instrument for the Brazilian population to measure personality under the FFM framework. Evidence of validity and reliability will be shown as well as results from DIF analysis.

(Invited Session) Where Do We Go from Here? A Practitioner's Discussion of Our Post-Pandemic World

1:00 to 2:00 pm - Organized Discussion

All of us were forced to change and adapt because of the global pandemic in a large number of ways including as professionals serving the educational system. We've seen assessments cancelled, test administration windows extended, test modes change, tests uses altered and the accuracy of interpretations complicated by various factors. The impacts of the pandemic will not only last until we return to our new normal, but will likely continue to resonate and impact the educational system into the future. Perhaps some changes implemented over the last year will remain indefinitely. As a result, it may not be tenable to continue to develop, administer, and score assessments in the same ways we have done in the past. Rather there are likely ways in which our industry will need to change and adapt in order to continue to positively contribute to education and to respond to the needs of students, educators, parents, and educational leaders in our post-pandemic world. This session features practitioners who will share their thoughts related to the ways in which their programs will change moving forward as well as the challenges that remain for our professional community.

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

Presenters:
Patrick Meyer, NWEA
Sharyn Rosenberg, NAGB
Melinda Ann Taylor, ACT

Electronic Board Session #2

1:00 to 2:00 pm

Participants:

A Comparative Study of Natural Language Processing Methods for Enemy Item Detection

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

This study explored the robustness of using Natural Language Processing for automatic enemy item detection. The performance of the Vector Space Model, the Latent Semantic Analysis, and the Latent Dirichlet Allocation was examined across multiple classifiers and probability cutoffs.

Administration of Selective Domain Tests in Benchmark Adaptive Testing

Rong Jin, Riverside Insights; Unhee Ju, Riverside Insights; JongPil Kim, Riverside Insights

The mathematics domains in a grade generally differ in difficulties and time of the academic year being taught to students. This study explores the effects of selective domain tests on ability estimates in CAT by comparing to the original full test.

An IRT Mixture Model for Item Position Effects

Klint Kanopka; Ben Domingue, Stanford University

Previous work on item position effects takes an item-side or person-side approach. We propose an IRT mixture model that estimates both, exploiting within-item variance in test position. This model is described, fit to a large item response dataset, and results are interpreted relative to published scores.

Applications of an exploratory sparse latent class model with polytomous attributes

Siqi He; Steven Culpepper, University of Illinois at Urbana-Champaign; Jeffrey Douglas

In this study, an exploratory sparse latent class model with polytomous attributes was proposed and applied to a big five personality test data. We are interested in discovering the item-attribute structure from a diagnostic model's perspective and comparing its model fit with the traditional item factor model.

A Sensemaking Approach to Understanding Teacher Interactions with Interim Assessments

Justin Paulsen, HumRRO

This study employs sensemaking approach to understanding how teachers respond to interim assessment (IA) use mandates. Using interview and survey data, this study examines how teachers make sense of IAs in one state's balanced assessment system. Findings describe how teacher perceptions and self-image can limit the instructional uses of IAs.

Assessing Differential Item Functioning (DIF) for a Large Number of Groups

Michelle Y. Chen, Paragon Testing Enterprises; Taylor Asbury, Paragon Testing Enterprises; Lok H. Chau, Paragon Testing Enterprises; William Tang, Paragon Testing Enterprises; Bruno D. Zumbo, University of British Columbia

Because most DIF detection methods were designed to compare two groups, DIF analysis remains a challenge when comparing a large number of groups. Using an example of DIF for a listening item attributable to test takers' first language (L1), we demonstrated a mixed-effects logistic regression method with 46 L1 groups.

Bayesian Item Response Theory Model Selection: Conditional and Marginal Likelihoods

Nnamdi Ezike, University of Arkansas; Allison Ames Boykin, University of Arkansas; Brian C Leventhal, James Madison University

Bayesian information criteria are computed on conditional or marginal likelihoods. This study uses simulation technique to assess relative model-fit of three dichotomous (1PLM, 2PLM, and 3PLM) and three polytomous (GRM, PCM, and GPCM) item response theory models based on conditional and marginal likelihoods of three information criteria.

Detecting Differential Item Functioning with Multiple Causes: A Comparison of Three Methods

Xiaowen Liu; H. Jane Rogers, University of Connecticut

Differential item functioning (DIF) is often caused by multiple sources. The effect of number of secondary dimensions on DIF detection rates was investigated by three DIF methods. Results showed that both the Mantel–Haenszel and logistic regression procedures performed well in detecting DIF in the presence of multiple secondary dimensions.

Do the Tenets of Answer Changing Research Hold for an Innovative Assessment?

Katherine Furgol Castellano, Educational Testing Service; Jamie Mikeska, ETS; Jung Aa Moon, Educational Testing Service; Steven Holtzman, Educational Testing Service; Jie Gao, Educational Testing Service; Yang Jiang, ETS

An analysis of pre-service teachers' answer changing behavior on an assessment of content knowledge for teaching (CKT) about matter and its interactions using technology-enhanced items further supports a long history of research showing the benefits of answer changing. Implications for test preparation, item development, and online assessment navigation are discussed.

High School Dropout Prediction Using Course-taking Patterns

Burhan Ogut, American Institutes for Research; Ruhan Circi, American Institutes for Research; Charles Scott, American Institutes for Research; Nevin Dizdari, American Institutes for Research

Current study leverages advances in predictive modeling to examine high-school dropout. We used student characteristics, course-taking, parental characteristics, and student expectations from the High School Longitudinal Study of 2009. Logistic regression, random forest, gradient boosting, and support vector machines were used and evaluated to predict dropping out of high school.

Investigate the Psychometric Properties of Multiple Forms based on Automatic Item Generation

Yishan Ding, University of Maryland; Jaehwa Choi, George Washington University; Hong Jiao, University of Maryland; Ji Hoon Ryoo, Yonsei University

This study utilizes empirical data to investigate the equivalence between test forms with items created using the Automatic Item Generation technique. It presents a multifaceted approach that includes construct equivalence evaluation, test-level comparison, and item-level comparison.

Masuring Strategic Diversity in Classrooms

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

Most research has focused on intraindividual strategy variability but far less known about inter-individual (or classroom) strategy diversity. This study presents a method to capture classroom-level strategic diversity, and we illustrate the formulation process with a research-based early mathematics assessment. The generated diversity scores have been applied to recent work.

Multilevel Measurement Invariance of the 2018-19 North Carolina Kindergarten Entry Assessment

Timothy Scott Holcomb; Richard Lambert, UNC Charlotte; Bryndle L Bottoms, University of North Carolina at Charlotte; Kawanna Jackson, UNC Charlotte

Most measurement invariance tests in education occur at the single level and ignore the nested structure of data. This study investigated an approach to test multilevel measurement invariance. The 2018-19 North Carolina Kindergarten Entry Assessment was used to test for measurement invariance across gender and home language.

Practical Utility of Proportion Agreement and Concordance Index for Item Fit

Insu Paek, Florida State University; Hirotaka Fukuhara, Pearson; Lanrong Li, Florida State University

To complement the existing well-known item fit statistical test procedures such as Q_1 and S-X^2, descriptive item fit measures which are easy to compute and easy to understand the degree of item fit in IRT applications are examined for their utilities, which are proportion agreement and concordance index.

Predicting Item Difficulty using Text Mining Techniques

Weimeng Wang, University of Maryland, College Park; Ray Yan, FINRA; Jeffrey Patton, FINRA

Due to practical or security reasons, field testing is not always a viable option for test developers. This study proposed to automate item difficulty prediction using text mining techniques. Overall, the models performed well and suggest they can be fruitfully applied to the domain of finance.

Preliminary Validity Evidence Supporting Direct Behavior Rating-Classroom Management (DBR-CM) Use in Secondary Settings

Elissa Mara Monteiro, University of California, Riverside; Nina Mandracchia, University of California, Riverside; Wesley Sims, University of California, Riverside

The Direct Behavior Rating-Classroom Management (DBR-CM) was developed to serve as a feasible, defensible, and flexible assessment of classroom management. This study continues the accumulation of evidence (i.e., concurrent validity and interrater reliability) in support of the stated interpretation and use argument (IUA) for DBR-CM.

ScratchWork Tool Usage and Its Relation to Performance in NAEP

Fusun Sahin, American Institutes for Research

Little is known about on-screen scratchworktool (SWT) use in digital assessments, e.g., whether it improves performance. We analyzed process data and scores from 2019 digitally-based NAEP mathematics. Results indicated that less than half of the students used SWT and SWT use was a significant contributor of performance in some items.

Score-level Sample Size Requirements for Technology-enhanced Items: A Simulation Study

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

The use of technology-enhanced items (TEIs) brings great possibilities for item development and item scoring. This study explores the impact of insufficient score-level sample size on item characteristics and data-model for TEIs that can be better interpreted by polytomous scoring with the use of partial credit model.

Subject Matter Experts' Judgments of Item Writing Difficulty Predict Item Characteristics

Rebecca Berenbon; Bridget McHugh, Center on Education and Training for Employment

SMEs estimated their ability to write test items to content standards. Average content standard ratings were used to predict item characteristics. SMEs' ratings were positively correlated with item difficulty. After controlling for item difficulty, higher ratings were associated with lower nonfunctioning distractor count but were not predictive of discrimination.

The Comparison between IRT Models and Generic Algorithms on Abbreviating Cognitive Tests

Yiling Cheng, Michigan State University

The aim of the present study is to compare the performances of shortening the cognitive tests between IRT diagnostic models and generic algorithm. To do so, the study is conducted with simulated data from IRT, DCM, and MIRT models, as well as a real data application.

The Golden and Silver Anchors: Scale Anchoring with Approximation Equations

Jonathan Rollins, West Virginia Department of Education

A new set of equations are presented that allow for approximate conversions between classical p-values and item response model (IRM) difficulty parameters. This research provides not only a way to identify IRM scales, but a way to approximately translate p-values to item difficulty parameters with a considerable degree of accuracy.

Using Graphical Model to Jointly Model Process data and Response Data

Jie Gao, Educational Testing Service; Matthew Johnson, ETS; Xiang Liu, Educational Testing Service

This paper tries to jointly model the process data and response data in writing assessment, using data from a middle school writing assessment. Using essay length as an example process feature, the model will evaluate the impact of the process feature on the effect of writing ability on item scores.

Two-tier Inference of Latent Traits and Cognitive Attributes

Hyeon-Ah Kang, University of Texas at Austin

The study proposes a two-tier modeling framework that allows simultaneous inference of latent traits and cognitive attributes. The framework is fit with full-information maximum likelihood and achieves greater precision in the inference of the latent factors than the higher-order model.

Understanding Test-Taking Cues: Experimental Findings on Examinee Score, Item Difficulty, and Format

Sarah Linnea Toton, Caveon Test Security; Tara Williams, Caveon

Savvy test-takers use test-taking cues to boost scores regardless of content knowledge. Cues are item characteristics that influence the likelihood of choosing a response, unrelated to item content. Experimental results show that examinee score, item difficulty, and item format impacted use of cues.

Writing Performance and Digital Familiarity: Multi-Group SEM Approach

Robert N Padgett; Young Yee Kim, American Institutes for Research; Xiaying Zheng, American Institutes for Research; Xiaoying Feng, American Institutes for Research

Using multi-group structural equation modeling, this study explores if digital familiarity measured by prior exposure to writing on a computer (PEWC) is related to writing performance (WP) and if the relationship varies across major subgroups (sex and race). The NAEP 2011 grade-8 writing digitally based assessment data were used.

A Bayesian Limited Information Approach to Diagnostic Classification Model-Data Fit

Catherine Elizabeth Mintz, University of Iowa; Jonathan Templin, University of Iowa; Jihong Zhang, University of Iowa

This study investigates an explicitly Bayesian PPMC method using a diagnostic classification model (DCM) as example. A Bayesian limited-information saturated model was fit and used as a comparison to a hypothesized DCM. Results suggest the Bayesian limited information saturated model approach is accurate and superior to traditional PPMC methods.

Going for Broke: Acknowledging and Disrupting the Barriers to Black Lives Mattering in Measurement

1:00 – 2:00 pm - Organized Discussion

We know that students - especially Black, Brown, and Indigenous students- do not experience the world (to include schooling) in ways that are context-free or color-neutral, so the question becomes why do we insist that they experience their assessments in this way? This question is asked in the middle of a global pandemic that has had a disproportionately devastating impact on communities of color (despite representing less than 15% of the population, Black Americans account for 25% of Covid cases and 39% of Covid-related deaths) accompanied by persistent racial injustices perpetrated in Black, Brown, and Indigenous communities/neighborhoods and schools.

Although many in our profession agree that an assessment design process that values and affirms the identities of minoritized students is critical, most have also focused on the barriers to implementation. These barriers include concerns about (a) costs; (b) the impact on non-White students; (c) potential trauma to marginalized students; (d) community buy-in; and (e) lack of representation in the field. As an extension of the Black Lives Matter in Educational Measurement organized discussion panel, this interactive session – which includes both small- and large-group discussions, will focus on how we can disrupt barriers to developing and maintaining assessment systems that are both culturally sustaining and antiracist.

Session Organizer:
Jennifer Randall, University of Massachusetts Amherst

Chair:
Taisha Steele, Pearson

Presenters:
Decca Knight
Pohai Kukea Shultz, University of Hawaii
Nirupa Matthew, Curriculum Associates
Maria Hamdani, Curriculum Associates
Kristen Huff, Curriculum Associates

(Invited Session) The Future of College Admissions Testing

2:15 to 3:45 pm - Organized Discussion

Long-term trends, recent scandals, legal challenges, as well as the impact of the pandemic, among other factors, have caused substantial challenges to the use of testing in college admissions. On the other hand, the need to obtain truthful evidence reflective of an applicant's achievement and potential for success in college has not suddenly disappeared. The rapid and sometimes radical changes beg questions and discussions about the future of testing in college admissions, and perhaps more importantly, about the potential mismatch between what have been in place versus what should ideally be developed and implemented vis-à-vis the next generation systems of assessments to better serve the needs of students and their families, K12 systems, and higher education institutions. Recent experience also suggests that in any such contemplation of the future of educational assessments in college admissions, one must address, in an open and constructive manner, the apparent gaps in performance among student subgroups, particularly along the lines of family wealth and race/ethnicity. This session brings together the perspectives from policy, psychometrics, higher education, and K12 education on the future of testing in college admissions in a moderated discussion format.

Session Organizer:

Li Cai, UCLA

Chair:

Patricia Gandara, Civil Rights Project at UCLA

Presenters:

Li Cai, UCLA
Sibyll Catalan, Geffen Academy at UCLA
Stephen Handel, College Board
Rebecca Zwick, Educational Testing Service

Rosetta Stone or Tower of Babel? Debating methods for NAEP-linked aggregate scores

2:15 to 3:45 pm - Organized Discussion

Educational researchers interested in comparing district-level achievement across states enter a metaphorical Tower of Babel. Differing tests and score scales appear to prevent cross-state analysis. Since the 1990s, the National Assessment of Educational Progress (NAEP) has tempted researchers as a possible Rosetta Stone, a decoder that might enable linking among state test score scales. In the widely cited Uncommon Measures, the 1999 National Research Council report, Feuer et al. seemed to rule decisively on NAEP linkages, "comparing the full array of… tests to one another… is not feasible" (p. 5). This symposium reopens this debate. A focal paper by Reardon, Kalogrides, and Ho (2021) proposes a method that evaluates aggregate linkages empirically. They show how NAEP district assessment scores can serve as a direct validation check. Based on their results, they conclude that linked estimates can support aggregate-level educational research. In a set of commentaries published with the focal paper in the Journal of Educational and Behavioral Statistics, five authors raise issues, cautions, and critiques of the method and uses of linked scores. Through short presentations and moderated debate, presenters will attempt to achieve consensus or draw clear lines of disagreement about how or whether to use NAEP-linked aggregate scores.

Session Organizer:

Andrew Ho, Harvard Graduate School of Education

Chair:

Steve Culpepper, University of Illinois

Presenters:

Daniel McCaffrey, Educational Testing Service
Andrew Ho, Harvard Graduate School of Education
Daniel Bolt, University of Wisconsin-Madison
Alina von Davier, Duolingo
Mark Davison, University of Minnesota
Tim Moses, College Board
Neil Dorans, Educational Testing Service
Sean Reardon, Stanford University

Impact of COVID-19 on Assessment

2:15 to 3:45 pm - Paper Session

Chair:

Aileen Reid, UNC Greensboro

Participants:

Assessment in the Time of Covid-19: Engagement During Low-Stakes Remote Testing

Steven Wise, NWEA; Megan Kuhfeld, NWEA; John Cronin, Northwest Evaluation Association

This study looked at the impact on test-taking engagement of moving from in-school interim achievement testing in fall 2019 to remote testing during spring 2020. Results showed that engagement was virtually unchanged and suggest that remote testing may not diminish student engagement if engagement features are provided.

Measuring Student Achievement in the Era of COVID-19

Qi Qin, Gwinnett County Public Schools; Shanna Ricketts, Emory University; Miranda McLaren, Gwinnett County Public Schools; James Appleton, Gwinnett County Public Schools

Gwinnett County Public Schools seeks to understand any disproportionate impact on student learning caused by COVID-19. The district has developed formative assessments for both in-person and digital students in SY2020-21. The opportunity to learn data collected by the district provides additional context on student achievement.

Multigroup sLDA for characterizing college students' research experience during COVID-19 pandemic

Hye-Jeong Choi, University of Georgia; Juyeong Lee, University of Georgia; Seohyun Kim, University of Virginia; Benjamin Listyg, University of Georgia; Brook Bowers, University of Georgia; Allan Cohen, University of Georgia; Erin Dolan, University of Georgia; Juan Ramírez-Lugo, University of Puerto Rico; Kyle Johnsen, University of Georgia

Ecological Momentary Assessment (EMA) is a longitudinal data collection method in which participants respond in real-time. Topic models are statistical models for extracting latent themes from a collection of documents. We apply a topic model to EMA text data for describing students' research experiences during COVID-19 pandemic.

Performance Based Assessment during COVID: Potential for Improving Learning through Crisis

Medjy Pierre-Louis; Elena Diaz-Bilello, University of Colorado Boulder, Center for Assessment, Design, Research and Evaluation (CADRE)

This paper will share findings from ongoing case studies taking place in two high schools using curriculum embedded performance-based assessments as a transformational tool to promote deeper learning and to evaluate whether students are receiving the skills needed to fulfill graduation requirements, particularly now in light of challenges presented by the ongoing health pandemic

Summer Slide Is Bad, COVID-19 Slide Is Even Worse: Online Assessment Perspective

Chalie Patarapichayatham, Southern Methodist University; Victoria Locke, Istation; Sean Lewis, Istation

This study investigates COVID-19 slide impacts on K-5 students' abilities in reading and math with 3-year online assessment data. Linear growth and estimated time score growth models are applied to estimate students' performance (before, during, and after COVID-19), summer slide (before and during COVID-19), COVID-19 slide, and no COVID-19.

Discussant:

Tracy Gardner, Classic Learning Test

Applications of Process Data

2:15 to 3:45 pm - Paper Session

Chair:

Leslie Keng, Center for Assessment

Participants:

Diagnosing Adult Information Problem Solving Strategy by Mining Clickstream Data

Yizhu Gao; Okan Bulut, University of Alberta; Ying Cui, University of Alberta

To compare strategies for well- and ill-defined information problem solving (IPS), we analyzed clickstream data of two tasks from the Programme for the International Assessment of Adult Competencies (PIAAC) 2012. Results revealed four consistent strategies and one unique strategy between two tasks, which differed in efficiency and effectiveness in IPS.

Measuring Item Process Data with Network Analysis Methods

Ni Bei, University of Washington; Elizabeth A. Sanders, University of Washington, Seattle; Nathan Abe; Min Li, University of Washington; Youngwon Kim, University of Washington

This paper proposes network analysis modeling as a novel approach for quantifying item process data at student, item, and student-item hybrid levels. We demonstrate these methods using 30 elementary students' action sequence times (eight per item) on two comparable math items that differed primarily in type of verbal information presented.

Latent Space Model for Process Data: An application of Partial Scoring

Yi Chen, Teachers College, Columbia University; Young-Sun Lee, Teachers College, Columbia University; Yi Yang, Columbia University; Jingru Zhang, Teachers College, Columbia University

The human-computer response process provides opportunities for extracting useful information on problem-solving. Zhu (2016) introduces social network methods for visualization. However, the research about modeling and extracting information from the response process with scoring are still limited. In this study, we extract latent positions of actions from the transition network using the latent space model. Then, we calculate the distance between different response sequences and rank the task-takers. The empirical data from PISA 2012 is discussed as a real example.

Modeling Changes in Response Style with Longitudinal IRTree Models

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

We propose a significant expansion to IRTrees to account for complex response processes across multiple time points using the item response tree framework. This study expands on a longitudinal IRTree, provides a simulation study to demonstrate adequate person- and item-parameter recovery in a Bayesian framework, and presents an empirical example.

Exploring Relationships Among Examinees' Behaviors and Performance Using Response Process Data

Sergio Araneda, University of Massachusetts Amherst; Burcu Arslan, Educational Testing Service; Madeleine Keehner, Educational Testing Service; Dukjae Lee, University of Massachusetts Amherst; Blair Lehman, Educational Testing Service; Jennifer L. Lewis, University of Massachusetts Amherst; Jung Aa Moon, Educational Testing Service; Stephen Sireci, University of Massachusetts Amherst

In this study, we analyzed students' process data (e.g., total time, first response latency) in a computer-based mathematics achievement test that included different item formats. The results indicated differences in response behaviors across item formats. Implications for test construction and future analyses in this area are discussed.

Discussant:

Tiago A. Caliço, American Institutes for Research

The AERA/APA/NCME Standards: Is it time to revisit the policy of self-enforcement?

2:15 to 3:45 pm - Organized Discussion

The Standards for Educational and Psychological Testing offer the most authoritative set of guidelines for professional practice in the field, but are explicitly designed without a formal framework for compliance, or other enforcement mechanisms. A lack of meaningful peer review, accountability, or even discussion is cause for concern in a context of continued growth but also mounting criticism of high stakes testing in education. This symposium brings together a group of leaders in the measurement community to discuss what NCME can do to increase compliance with the Standards and help bring research and professional practice in the field in closer alignment. Critical issues include among others, how to decide when violations of the Standards rise to a particular level of concern and what actions are appropriate in these cases; what internal or independent structures could give the Standards more teeth; what is the ideal composition of these bodies and show should they balance different stakeholder perspectives; and how make explicit technical and substantive concerns related to equity. The audience will be able to register questions, contributions, and commentary—the session ultimately aims to start a conversation among the NCME membership, and delineate an agenda for formal discussion and action.

Session Organizer:

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

Presenters:

Drew Gitomer, Rutgers University
Joan Herman, UCLA CRESST
Daniel Koretz, Harvard Graduate School of Education
Scott Marion, Center for Assessment
Stephen Sireci, University of Massachusetts Amherst

Fairness for All Examinees

2:15 to 3:45 pm - Paper Session

Chair:

Dubravka Svetina Valdivia, Indiana University

Participants:

Timing Analyses for Exploring Racial Equity in Testing

Dukjae Lee, University of Massachusetts Amherst; Carol Ezzelle, National Board for Professional Teaching Standards

This research explored the differences in item response time between Blacks/Hispanics and Whites with similar proficiency levels. Two-way ANOVA yielded no significant difference in item response time of the two ethnic groups. Results did not imply racial biases, and further practical considerations are discussed.

Using Examinee Characteristics to Predict Individual Differences in Item Position Effects

Thai Quang Ong, National Board of Medical Examiners; Dena Pastor, James Madison University

We explored the relationships between six examinee characteristics and item position effects in two low-stakes tests. Gender, change in fatigue, and change in boredom were significant Chairs of item position effects; however, the magnitude and significance of these effects differed across the two low-stakes tests.

Should psychometricians make claims about fairness?

Daniel Katz, University of California, Santa Barbara; Anthony Clairmont, University of California, Santa Barbara

The Standards for Educational and Psychological Testing (AERA, NCME, APA, 2014) have a chapter on test fairness that neglects to define fairness. We present common definitions of fairness and question whether these definitions of fairness are probed by psychometrics. We map test results across New York City as an example.

An equity-based approach to developing and validating culturally-sensitive assessments

Stanley N Rabinowitz, Pearson

We describe how cultural- and linguistic-based meaning-making impact assessment outcomes and the validity of interpretations across student subgroups. We demonstrate how a coherent, equitable and culturally sensitive assessment system can address implicit and explicit biases of the several stakeholders, resulting in fair accountability and instructional strategies tailored to each student.

Fairness in assessment - contemporary challenges and implications for practice

Isabel Nisbet, University of Cambridge; Stuart Shaw, Cambridge International

This presentation will question current theoretical understanding of fairness and address three contemporary challenges to fair assessment relating to:

- developing views of validity
- views of social justice in terms of 'educational adequacy'
- 'Big Data' - which poses new questions about validity and reliability, as well as fairness.

Discussant:

Kristen Huff, Curriculum Associates

Security Issues in Education

2:15 to 3:45 pm - Paper Session

Chair:

Rich Feinberg, National Board of Medical Examiners

Participants:

Does Item Format Affect Test Security?

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

Unlike the traditional multiple-choice (MC) format, the discrete-option multiple-choice (DOMC) format does not necessarily reveal all answer options to an examinee. This study considers whether the reduced exposure of item content affects test security.

Online Detection of Aberrant Test-taking Behaviors

Suhwa Han, University of Texas at Austin; Hyeon-Ah Kang, University of Texas at Austin

The study proposes sequential monitoring procedures that can surveil examinee behaviors in computer interactive assessments. The procedures will monitor examinees' responses and reaction times, and signal aberrancy in real time. The study investigates adequacy of the procedures in simulations and two empirical settings: computer-based testing and online learning environments.

Which Statistics Should be Used Under the Uncertainty in Compromised Item Set?

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

This study proposes two new statistics relying on responses and response times to detect item preknowledge. The performance of the statistics is investigated when there is uncertainty in the set of suspicious items for both adaptive and linear tests through a simulation study and a real data application.

Application of l! statistic in identifying aberrant patterns in nominal responses

Tianpeng Ye; Louis Roussos, Cognia; Xi Wang, Cognia; Liuhan Cai, Cognia

Kim and Roussos (2017) presented a case of large violations of local independence related to how students selected options on multiple-choice items. This study extends the l! statistic to the nominal response data to see if it can give a deeper understanding and improved detection of this aberrant response behavior.

A New Method of Detecting Cheating due to Item Compromise

Lei Wan; Weiwei Cui, College Board

This paper presents a new approach to detecting cheating when a subset of test items was leaked, and some students had pre-knowledge of the leaked items. A simulation study was conducted to evaluate the usefulness of the approach and to compare it with the existing Likelihood Ratio Test method.

Discussant:

Matthew N. Gaertner, WestEd

Exploring Scoring and Psychometric Modeling Strategies for Innovative Items

4:00 to 5:30 pm - Coordinated Paper Session

As assessments have transitioned from paper-and-pencil formats to digital formats, testing programs have been increasingly including innovative or technology enhanced (TE) items for a variety of purposes. These items can be classified across several dimensions of innovation, including: item format, response action, media inclusion, level of interactivity, and scoring method (Parshall, Davey, & Pashley, 2000). This session explores various ways in which the first four dimensions may influence the scoring method dimension, defined as the method by which examinee responses are translated into quantitative scores (Parshall et al., 2000). The scoring dimension incorporates both scoring rules to determine the number and order of score levels for innovative items and the modeling strategy employed. Each of the papers in this session includes analysis of real data, which together encompass a wide variety of innovative item types from several testing programs. The session includes commentary from a leading researcher in psychometric methods.

Session Organizer:

Carol Eckerly, Educational Testing Service

Chair:

Adrienne Sgammato, ETS

Participants:

A Framework for Rule-Based Scoring of Technology Enhanced Items

William A. Lorie, Center for Assessment

Exploring Strategies for Optimal Scoring Rubric Development of Technology Enhanced Items

Adrienne Sgammato, ETS

Item Response Theory Models for Adaptive Testlet Items

Carol Eckerly, Educational Testing Service; Paul Adrian Jewsbury, Educational Testing Service; Yue Jia, Educational Testing Service

A Comparison of Rapid Picture and Rapid Color Naming Screeners

Adam Wyse, Renaissance; Scott R. McConnell, Renaissance and University of Minnesota; Eric Stickney, Renaissance; Catherine Close, Renaissance Learning; Heidi Lund, Renaissance

Evaluating Bias from Introducing New Items Into a Scale

Paul Adrian Jewsbury, Educational Testing Service; Ru Lu, Educational Testing Service

Discussant:

Billy Skorupski, Questar Assessment

IRTree Models: The Illus-tree-ous and In-tree-guing Response Process Models

4:00 to 5:30 pm - Coordinated Paper Session

IRTrees encompass a multi-dimensional item response theory family of models that explicitly model response processes. These tree-like models have recently garnered significant research attention due to their vast number of diverse applications combined with the wide-spread availability of advanced computing power. The five presentations in this session will exhibit the versatility and utility of the IRTree family of models through empirical applications and methodological innovation research. Presentation 1 will present a novel approach to incorporating response style information from anchoring vignettes in order to estimate a trait of interest from self-report items. Presentation 2 will combined a hierarchical speed and IRTree model for investigating response processes. The third presentation will explore the validity of IRTree inferences through factor analysis, simulation, and the calculation of pseudo-item information. The fourth presentation will show the diversity of applications IRTrees can be applied to by estimating an ability when two-attempt multiple choice items are used. And, the fifth presentation will assess the stability of response styles by employing an IRTree model across independent traits. The five presentations will jointly display the flexibility and usefulness of the IRTree response process models.

Session Organizer:

Brian C Leventhal, James Madison University

Participants:

A tree-based approach to identifying response styles with anchoring vignettes

Brian C Leventhal, James Madison University; Christina K Zigler, Duke University School of Medicine

A Combined Hierarchical Speed and IRTree Model for Investigating Response Processes

Aaron Myers, University of Arkansas; Allison Ames Boykin, University of Arkansas

Validity Evidence for Response Process Models

Allison Ames Boykin, University of Arkansas; Aaron Myers, University of Arkansas

Modeling Two-Attempt Multiple-Choice Items Using IRTrees

Philip Grosse, University of Pittsburgh; Clement Stone, University of Pittsburgh

An IRTree method to investigating stability of extreme and midpoint response style

Nikole Gregg, Cambium Assessment, Inc.; Brian C Leventhal, James Madison University

Innovations in Detection of Test Collusion

4:00 to 5:30 pm - Coordinated Paper Session

This coordinated session aims to help close the gap between research and practice in detection of compromised groups, which remains as one of the most significant challenges in test security. This session brings together several leading scholars in the area of test security to introduce and discuss four novel approaches to detecting collusion. Collectively, the set of papers will introduce new approaches which simultaneously detect compromised items and the groups of individuals engaged in collusion, leverage the latest technology and computational methods to improve detection through machine learning, sequential clustering, and the incorporation of eye tracking process data, and investigate the efficacy and impact of real-time data forensics to detect examinees with preknowledge during the exam.

Session Organizer:

James Wollack, University of Wisconsin

Participants:

Detecting Preknowledge Via Joint Modeling of Responses, Response Times, and Visual-Fixation Counts

Kaiwen Man, University of Alabama; Jeffrey Harring, University of Maryland; Youn-Jeng Choi, University of Alabama; James Wollack, University of Wisconsin

An Iterative Unsupervised-Learning-Based Approach for Detecting Item Preknowledge

Yiqin Pan, University of Wisconsin-Madison; James Wollack, University of Wisconsin

A New Approach to Detection of Collusion: Iterative Cluster Building

James Wollack, University of Wisconsin; Sakine Gocer Sahin, WIDA at UW-Madison

Reducing Score Bias Through Real-Time Rerouting of Examinees with Anomalous Responses

Merve Sarac, University of Wisconsin-Madison; James Wollack, University of Wisconsin

Discussant:

Cengiz Zopluoglu, University of Oregon

Moving Bookmark Standards Setting from In-Person to Virtual: Best Practices/Lessons Learned

4:00 to 5:30 pm - Organized Discussion

This session will discuss the procedures used to set standards on the Kansas English Language Proficiency Assessment (KELPA) in a virtual setting in fall 2020. KELPA serves as an excellent litmus test for other programs considering virtual standard setting as a potential alternative to in-person event due to the complex nature of the assessment. The session includes a brief overview of the processes and tools used for the virtual KELPA standards setting, followed by a collaborative discussion with audience members who may be considering conducting virtual standards settings. The session will attend to the challenges associated with moving the process to a virtual environment and how those challenges were overcome. All aspects of standards setting will be addressed, including: training and scripting for facilitators; creating and delivering online training for panelists; dealing with secure test materials in a virtual environment; conducting the workshop virtually; dealing with process documentation (evaluations and other logistical materials), utilizing content expertise during the panel meetings; and vertically articulating the ensuing results. The discussion panel will include participants from all phases of the process to answer and brainstorm standards setting solutions with audience members.

Session Organizer:
Arthur Thacker, HumRRO

Presenters:
Andrea Word, University of Alabama in Huntsville
Andrea Sinclair, HumRRO
Brooke Nash, University of Kansas
Jie Chen, University of Kansas

E ≠ MC2: Equity doesn't equal measurement calibration squared

4:00 to 5:30 pm - Organized Discussion

It's a satirical session. With short skits and parody commercials. Some mixed media if we get our act together. Otherwise, it's a variety show. It'll be fun. Maybe. Research shows almost half (46.7%) don't know or have no opinion if these are fun (Ackerman, 2015, p. 20).

Session Organizer:

Pamela Paek

Presenters:

Pamela Paek
Arthur Stanley, Pearson
Jennifer Randall, University of Massachusetts
Karla Egan, EdMetric, LLC
Chad Buckendahl, ACS Ventures
** with some special guest stars/cameos*

Topics in Linking and Equating

4:00 to 5:30 pm - Paper Session

Chair:

Jaime Malatesta, Graduate Management Admission Council

Participants:

Impact of item drift on ability estimation under LOFT and linear forms

Irina Grabovsky, National Board of Medical Examiners; Chunyan Liu, National Board of Medical Examiners; Raja G Subhiyah, National Board of Medical Examiners

Continuous administration of linear forms over long period of time may cause drift in item difficulty, which may in turn affect estimation of proficiencies. This research compares linear forms and LOFT models from the perspectives of precision of ability estimation and test security characteristics.

Anchor Item Replacement in the Presence of Consequential Item Parameter Drift

Kuo-Feng Chang; Kelly Rewley, American Board of Internal Medicine

When a high percentage of items are de-anchored for drift, the linking function can become unstable. One way to attempt to mitigate this issue is to engage in anchor item replacement in the linking process. In this study, we proposed and examined several anchor item replacement procedures using simulated data.

Comparing the Performance of Item Parameter Drift Detection Methods

Kelly Rewley, American Board of Internal Medicine; Pamela Kaliski, ABIM

The purpose of this simulation study was to compare the performance of four item parameter drift detection approaches. We crossed two methods (a-a/b-b plots, D2 index) with two cutoff criteria types (absolute, relative). Using the anchor composition index as our outcome, D2 with a 2 SD (relative) cutoff performed best.

Application of Matrix Completion Methods to Item Calibration in Target Testing Design

Yawei Shen; Shiyu Wang, University of Georgia

This research study investigates the possibility of applying a matrix completion method, a singular value thresholding (SVT) algorithm, in handling missing responses from targeted testing design. Simulation studies are conducted to investigate factors that impact their performance. Results demonstrate the potentials of applying the modified SVT in handling missing data.

Is Equating Necessary When Replacing a Small Number of Items?

Joshua MacInnes, Ascend Learning

This simulation study examined whether equating is necessary when replacing a small of number of items on a certification test. The results indicated that equating may not be necessary when replacing a single item or when replacing a small number of items with items of similar difficulty.

Discussant:

Anna Topczewski, WestEd

Topics in Cognitive Diagnosis Modeling

4:00 to 5:30 pm - Paper Session

Chair:

Melissa L. Gholson, Educational Testing Service

Participants:

A General Cognitive Diagnosis Model for Polytomous Attributes

Jimmy de la Torre, University of Hong Kong; Xue-Lan QIU, University of Hong Kong; Kevin Carl Santos, University of the Philippines

The pG-DINA model for polytomous attributes can benefit classroom instruction. However, it depends on a stringent assumption. This study proposes a general CDM that relaxes the assumption and allows different attribute levels to contribute differentially to the success probability. Simulation studies were conducted to evaluate the new model's parameter recovery.

Diagnostic Concept Inventories for Misconception Classification Accuracy and Reliability

Madeline Schellman; Laine Bradshaw, University of Georgia

We examined the diagnostic classification model (DCM) framework for providing accurate and reliable classifications of student misconceptions from diagnostic concept inventories. We investigated estimation under a variety of conditions based on the a priori design of a middle grades diagnostic concept inventory measuring probabilistic reasoning misconceptions.

Nonparametric Classification Method for Multiple-Choice Items in Cognitive Diagnosis

Yu Wang, Rutgers University; Chia-Yi Chiu, Rutgers University

A nonparametric classification method for multiple-choice items (MC-NPC) for cognitive diagnosis (CD) is proposed in the study. The preliminary simulation study shows that the MC-NPC method results in higher correct classification rates than the traditional CD methods for dichotomous data and outperforms the MC-DINA model when the samples are small.

Nonparametric Attribute Classification based on Saturated Cognitive Diagnostic Models

Sook Hyun Park; Hyeon-Ah Kang, University of Texas at Austin

The study proposes a general nonparametric classification framework for dichotomous and polytomous response data. The framework is developed within the saturated cognitive diagnostic model and can be applied to any item-attribute interaction. The study evaluates performance of the method in simulation and real data analysis.

A General Method for Q-Matrix Estimation

Jiaxi Wang, Rutgers University; Chia-Yi Chiu, Rutgers University; Hans Friedrich Koehn

The proposed Two-Step Q-matrix Estimation (TSQE) method uses factor analysis to construct a provisional Q-matrix which then is finalized with a Q-matrix renement/validation method. The TSQE method can be used with any cognitive diagnosis model and is computationally very economical.

Discussant:

Leah Feuerstahler, Fordham University

Conference Participants

Abulela, Mohammed, mhady001@umn.edu
Agard, Christopher, Educational Testing Service, cagard@ets.org
Ahumada, Audra, Arizona Department of Education, Audra.Ahumada@azed.gov
Albanese, Mark, National Conference of Bar Examiners, malbanese@ncbex.org
Albano, Anthony, University of California, Davis, adalbano@ucdavis.edu
Ali, Usama, Educational Testing Service, uali@ets.org
Alpizar, David, d.martinezalpizar@wsu.edu
Alves, Cecilia, Medical Council of Canada, calves@mcc.ca
Andersen, Nico, DIPF | Leibniz Institute for Research and Information in Education, Andersen.Nico@dipf.de
Araneda, Sergio, University of Massachusetts Amherst, saraneda@umass.edu
Arce, Alvaro J., Pearson, alvaro.arce-ferrer@pearson.com
Arena, Eric, Putnam County School District, eric_arena@putnam.k12.ga.us
Armour-Thomas, Eleanor, Queens College of the City University of New York, Eleanor.Armour-Thomas@qc.cuny.edu
Arslan, Burcu, Educational Testing Service, barslan@ets.org
Arthur, Ann, ACT, ann.arthur@act.org
Attali, Yigal, Duolingo, yattali@gmail.com
Awwal, Nafisa, University of Melbourne, n.awwal@unimelb.edu.au
Bailey, Alison, abailey@gseis.ucla.edu
Baker, Eva, UCLA, baker@cse.ucla.edu
Baker, Ryan, University of Pennsylvania, rybaker@upenn.edu
Bandalos, Deborah, James Madison University, bandaldl@jmu.edu
Bao, Yu, James Madison University, bao2yx@jmu.edu
Barragan Torres, Mariana, UCLA, marianabarragan@ucla.edu
Barrett, Michelle Derbenwick, Edmentum, mldbarrett@gmail.com
Barton, Karen, NWEA, karen.barton@nwea.org
Beard, Jonathan, College Board, jonathan.j.beard@gmail.com
Beck, Michael, BETA, LLC, mikebeck@prodigy.net
Becker, Kirk, Pearson, kirk.becker@pearson.com
Beckler, Amanda, Renaissance, Amanda.Beckler@renaissance.com
Behrendt, Stefan, University of Stuttgart, stefan.behrendt@ife.uni-stuttgart.de
Bei, Ni, University of Washington, nbei@uw.edu
Beimers, Jennifer, Pearson, jennifer.beimers@pearson.com
Beiting-Parrish, Magdalen, CUNY Graduate Center, mbeiting@gradcenter.cuny.edu
Bennett, Randy, ETS, rbennett@ets.org
Berenbon, Rebecca, berenbon.1@osu.edu
Betebenner, Damian, Center for Assessment, dbetebenner@nciea.org
Betts, Joseph, National Council of State Boards of Nursing, jbetts5118@aol.com
Beverly, Tanesia, Law School Admission Council, trbeverly@lsac.org
Bialo, Jacquelyn A., jbialo1@student.gsu.edu
Bian, Meina, University of Georgia, adele.bein@gmail.com
Bishop, Kyoungwon, WIDA at UW-Madison, kei.bishop2@gmail.com

Bolender, Brad, ACT, brad.bolender@gmail.com

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

Botha, Sandra Margaret, sbotha@umass.edu

Botter, Preston, UCLA CRESST, pdbotter@gmail.com

Boyer, Michelle, Center for Assessment, mboyer@nciea.org

Boykin, Allison Ames, University of Arkansas, boykin@uark.edu

Bradshaw, Laine, University of Georgia, laineb@uga.edu

Brandt, Chris, Center for Assessment, cbrandt@nciea.org

Breidenbach, Daniel H., PSI, dbreidenbach@psionline.com

Brennan, Robert, University of Iowa, robert-brennan@uiowa.edu

Brice, Amanda, Curriculum Associates, abrice@cainc.com

Bridgeman, Brent, ETS, bbridgeman@ets.org

Briggs, Derek, University of Colorado, derek.briggs@colorado.edu

Broer, Markus, American Institutes for Research, markus.broer@gmail.com

Brookhart, Susan, Duquesne University, suebrookhart@gmail.com

Browne, Mary, National Board of Certification and Recertification for Nurse Anesthetists, mbrowne@nbcrna.com

Brucia, Robert, American Board of Pediatrics, rcb0424@gmail.com

Buchanan, William Robert, SAG Corp, william@williambuchanan.net

Buckendahl, Chad W., ACS Ventures, LLC, cbuckendahl@acsventures.com

Bulut, Hatice Cigdem, Cukurova University, hcyavuz@cu.edu.tr

Bulut, Okan, University of Alberta, bulut@ualberta.ca

Burstein, Jill, Educational Testing Service, jburstein@ets.org

Cai, Li, UCLA, lcai@ucla.edu

Caliço, Tiago A., American Institutes for Research, tcalico@protonmail.com

Camara, Wayne J., LSAC, waynecamara@gmail.com

Cancado, Luciana, Curriculum Associates, lcancado@cainc.com

Cardwell, Ramsey, rlcardwe@uncg.edu

Carroll, Sarah, National Board for Certification in Occupational Therapy, scarroll@nbcot.org

Catalan, Sibyll, Geffen Academy at UCLA, scatalan@geffenacademy.ucla.edu

Cerentini Pacico, Juliana, University of Iowa, juliana-pacico@uiowa.edu

Chai, Emily, zchai9@gatech.edu

Chang, Hua-Hua, Purdue University, chang606@purdue.edu

Chang, Kuo-Feng, kuo-feng-chang@uiowa.edu

Chatterji, Madhabi, Teachers College, Columbia University, Chatterji@Exchange.tc.columbia.edu

Chen, Dandan, The American Board of Anesthesiology, chendan@udel.edu

Chen, Dandan, University of Illinois at Urbana-Champaign, dandan.c.chen@gmail.com

Chen, Jie, University of Kansas, xiaojiewd@hotmail.com

Chen, Jihang, Boston College, jihang@bc.edu

Chen, Jing, NWEA, jing.chen@nwea.org

Chen, Juan, National Conference of Bar Examiners, jchen@ncbex.org

Chen, Lida, lida-chen@uiowa.edu

Chen, Michelle Y., Paragon Testing Enterprises, mchen@paragontesting.ca

Chen, Wenya, Loyola University Chicago, wchen7@luc.edu
Chen, Yi, Teachers College, Columbia University, yc3356@columbia.edu
Cheng, Yiling, Michigan State University, yiling.cheng2@gmail.com
Chesluk, Ben, American Board of Internal Medicine, bchesluk@abim.org
Cho, YoungWoo, ACT, youngwoo.cho@act.org
Choe, Edison M., Graduate Management Admission Council, echoe@gmac.com
Choi, Hye-Jeong, University of Georgia, hjchoi1@uga.edu
Choi, Jinah, Jinah.Choi@edmentum.com
Choi, Jinnie, Savvas Learning Company, jinnie.choi@gmail.com
Christensen, Laurene, WIDA at the Wisconsin Center on Education Research, laurene.christensen@wisc.edu
Chung, Jinmin, University of Iowa, jinmin-chung@uiowa.edu
Cintron, Dakota Wayne, University of Connecticut, dakots3122@gmail.com
Circi, Ruhan, American Institutes for Research, rcirci@air.org
Clauser, Amanda, National Board of Medical Examiners, aclauser@nbme.org
Cleveland, Christopher, Harvard University, chcleveland@g.harvard.edu
Close, Kevin, Arizona State University, kclose1@asu.edu
Coggeshall, Whitney, American Board of Internal Medicine, wcoggeshall@abim.org
Cohen, Yoav, National Institute for Testing, coyoav@gmail.com
Cong, Menglong, University of Denver, the.menglongcong@gmail.com
Conley, David, EdImagine, David_Conley@edimagine.com
Conrad, Zachary, USD 497, zachary.conrad@usd497.org
Cook, Robert, Cognia, robert.cook@cognia.org
Coppens, Lindsay, Ontario Institution for Studies in Education, lindsay.coppens@mail.utoronto.ca
Cui, Weiwei, College Board, wcui@collegeboard.org
Cui, Zhongmin, CFA Institute, zhongmin.cui@cfainstitute.org
Culpepper, Steven, University of Illinois at Urbana-Champaign, sculpepp@illinois.edu
Daisher, Ted, University of Massachusetts Amherst, tdaisher@umass.edu
Dallas, Andrew, National Commission on Certification of Physician Assistants, drewd@nccpa.net
Davenport, Ernest, University of Minnesota, LQR6576@umn.edu
Davis, Laurie, Curriculum Associates, laurie@davistx.com
Davis-Becker, Susan, ACS Ventures, LLC, sdavisbecker@acsventures.com
Dawber, Teresa, Council for Aid to Education, tess.dawber@att.net
de la Torre, Jimmy, University of Hong Kong, j.delatorre@hku.hk
Demir, Cihan, cihan.demir@wsu.edu
Demirkaya, Onur, University of Illinois at Urbana-Champaign, onurdmrkaya@gmail.com
Denbleyker, John, Houghton Mifflin Harcourt, jdenblge@yahoo.com
Deng, Jiayi, jiayideng0726@gmail.com
Deng, Nina, Kaplan INC., nndeng@gmail.com
DiCerbo, Kristen, Pearson, kdicerbo@cox.net
Dickinson, Emily, HumRRO, edickinson@humrro.org
Dimitrov, Dimiter Milkov, National Center for Assessment, ddimitro@gmu.edu
Ding, Yishan, University of Maryland, ysding@umd.edu
Ding, Zhaopeng, dingzhaopeng2013@gmail.com

Doe, John, mgaertn@wested.org
Dolan, Robert, Diverse Learners Consulting, rdolan@alum.mit.edu
Domaleski, Chris, Center for Assessment, cdomaleski@nciea.org
Dong, Yixiao, University of Denver, yixiao.dong@du.edu
Donoghue, John, Educational Testing Service, jrdonoghue@comcast.net
Dosedel, Michael, University of Minnesota, dose0018@umn.edu
Dray, Amy, Spencer Foundation, ajdray@gmail.com
Drost, Bryan R., Rocky River Schools, drostbr@gmail.com
Du, Yang, yangdu015@gmail.com
Du, Ying, American Board of Pediatrics, ydu@abpeds.org
Dumas, Denis, University of Denver, denis.dumas@du.edu
Dumoulin, Amanda Rose, Kwantlen Polytechnic University, amanda.dumoulin@kpu.ca
Dunn, Jennifer, Questar, jdunn@questarai.com
Durrence, Debbie, Gwinnett County Public Schools, debbie.durrence@gcpsk12.org
Dwyer, Andrew, American Board of Pediatrics, adwyer@abpeds.org
Early, Kellie, National Conference of Bar Examiners, kearly@ncbex.org
Eckerly, Carol, Educational Testing Service, ceckerly@ets.org
Englert, Kerry, Seneca Consulting, LLC, kenglert@comcast.net
Ercikan, Kadriye, Educational Testing Service, kercikan@ets.org
Ersan, Ozge, University of Minnesota Twin Cities, ersan001@umn.edu
Evans, Carla M., cevans@nciea.org
Ezike, Nnamdi, University of Arkansas, ncezike@uark.edu
Ezzelle, Carol, National Board for Professional Teaching Standards, cezzelle@nbpts.org
Fager, Meghan, National University, mfager@nu.edu
Fan, Fen, fenf@nccpa.net
Fechter, Tia, Office of People Analytics, tiacorliss@hotmail.com
Feinberg, Rich, National Board of Medical Examiners, rfeinberg@nbme.org
Feng, Gary, Educational Testing Service, gfeng@ets.org
Ferrara, Steve, Cognia, sferrara1951@gmail.com
Feuerstahler, Leah, Fordham University, lfeuerstahler@fordham.edu
Finch, Holmes, Ball State University, whfinch@bsu.edu
Firoozi, Tahereh, tahereh.firoozi@ualberta.ca
Flanagan, Kathleen, Massachusetts Department of Education, Kathleen.R.Flanagan@mass.gov
Flores, Charity, Indiana Department of Education, cflores@doe.in.gov
Fong, Karen, kfong6@uic.edu
Font, Maureen, *University of Illinois – Chicago,* maureenfont@gmail.com
Forte, Ellen, edCount, LLC, eforte@edcount.com
Foster, David, Caveon Test Security, david.foster@caveon.com
Francis, Catherine Xueying, Houghton Mifflin Harcourt, catherinehfrancis@gmail.com
Fu, Yanyan, GMAC, frankyanyan@gmail.com
Furgol Castellano, Katherine, Educational Testing Service, KEcastellano@ets.org
Furter, Robert Thomas, American Board of Pediatrics, rfurter@abpeds.org
Gaertner, Matthew N., WestEd, matt.gaertner@gmail.com

Gandara, Patricia, Civil Rights Project at UCLA, pcgandara@gmail.com

Gao, Jie, Educational Testing Service, jgao@ets.org

Gao, Yizhu, yizhu@ualberta.ca

Garcia, Elda, National Association of Testing Professionals, elda.garcia@natponline.com

Gardner, Tracy, Classic Learning Test, tgardner@cltexam.com

Ge, Yuan, University of Alabama, yge4@crimson.ua.edu

Geisinger, Kurt, University of Nebraska-Lincoln, kgeisinger2@unl.edu

Gezer, Tuba, tgezer@uncc.edu

Gholson, Melissa L., Educational Testing Service, mgholson@ets.org

Gianopulos, Garron, NWEA, garron@gianopulos.com

Gitomer, Drew, Rutgers University, drew.gitomer@gse.rutgers.edu

Gochyyev, Perman, University of California, Berkeley, perman@berkeley.edu

Goldhammer, Frank, DIPF | Leibniz Institute for Research and Information in Education, Centre f. Int. Student Assessm, goldhammer@dipf.de

Gong, Brian, Center for Assessment, bgong@nciea.org

Gong, Tao, Educational Testing Service, tgong@ets.org

Gonzalez, Jorge, Pontificia Universidad Catolic, jorge.gonzalez@mat.uc.cl

Gonzalez-Wegener, Xaviera, UCL Institute of Education, xgonzalezwe@gmail.com

Gordon, Edmund, John M. Musser Professor of Psychology, Emeritus - Yale University / Richard March Hoe Professor of, egordon@tc.edu

Gorgun, Guher, University of Alberta, gorgun@ualberta.ca

Gorham, Jerry L., Ascend Learning, jerry.gorham@gmail.com

Gorney, Kylie N., University of Wisconsin-Madison, kyliengorney@gmail.com

Gotzmann, Andrea Julie, Medical Council of Canada, agotzmann@mcc.ca

Gough, Michelle, EdMetric, LLC, michelle.r.gough@gmail.com

Grabovsky, Irina, National Board of Medical Examiners, igrabovsky@nbme.org

Gregg, Nikole, Cambium Assessment, Inc., Greggnl@jmu.edu

Grosse, Philip, University of Pittsburgh, pjg21@pitt.edu

Gueorguieva, Johnna, johnnag10@gmail.com

Guo, Hongwen, Educational Testing Service, hguo@ets.org

Guo, Wenjing, University of Alabama, wguo9@crimson.ua.edu

Guzman-Orth, Danielle, Educational Testing Service, dguzman-orth@ets.org

Hahnel, Carolin, DIPF | Leibniz Institute for Research and Information in Education, Centre for International Student, hahnel@dipf.de

Hajek, Andrea, National Board of Professional Teaching Standards, AHajek@nbpts.org

Hamdani, Maria, Curriculum Associates, mhamdani@cainc.com

Han, Suhwa, University of Texas at Austin, suhwa@utexas.edu

Handel, Stephen, College Board, shandel@collegeboard.org

Hansen, Mark, UCLA, markhansen@ucla.edu

Hao, Jia, University of Minnesota Twin Cities, jiaxx052@umn.edu

Hao, Jiangang, Educational Testing Service, jhao@ets.org

Harris, Deborah, University of Iowa, deborah-harris@uiowa.edu

Harris, William, Assoc. Of Test Publishers, wgharris@testpublishers.org

Hazen, Tim, Iowa Testing Programs, timothy-hazen@uiowa.edu

He, Qiwei, Educational Testing Service, qhe@ets.org
He, Siqi, siqihe2@illinois.edu
He, Wei, NWEA, heweit@gmail.com
He, Yong, ACT, yong.he@ceciic.org
Hembry, Tracey, Alpine Testing Solutions, Inc., tracey.hembry@alpinetesting.com
Hennessy, Briana, University of Connecticut, Briana.Hennessy@uconn.edu
Henson, Robert, University of North Carolina, rahenson@uncg.edu
Herman, Joan, UCLA CRESST, herman@cse.ucla.edu
Hess, Brian J, College of Family Physicians of Canada, bhess@cfpc.ca
Hicks, Juanita, American Institutes for Research, juanita.hicks.11@gmail.com
Himelfarb, Igor, ihimelfarb@nbce.org
Ho, Andrew, Harvard Graduate School of Education, Andrew_Ho@gse.harvard.edu
Holcomb, Timothy Scott, tholcom4@uncc.edu
Hong, Minju, University of Georgia, mh19985@uga.edu
Hong, Seong Eun, shong@umass.edu
Hood, Stafford, University of Illinois at Urbana-Champaign, stafford.hood@asu.edu
Hoover, Jeffrey, University of Kansas, jhoover4@ku.edu
Hu, Mingqi, University of Illinois at Urbana-Champaign, mh2@illinois.edu
Hua, Cheng, chua@crimson.ua.edu
Huang, Katherine, yueh@udel.edu
Huang, Qi, University of Wisconsin-Madison, qhuang85@wisc.edu
Huff, Kristen, Curriculum Associates, khuff@cainc.com
Huh, Nooree, ACT, nooreehuh@gmail.com
Hyslop, Alisha, Association for Career and Technical Education, ahyslop@acteonline.org
Ickes, Kelly, Cognia, kelly.ickes@cognia.org
Ihlenfeldt, Samuel Dale, University of Minnesota, ihlen010@umn.edu
Im, Sukkeun, NWEA, sukkeun@gmail.com
Ingrisone, Soo, Pearson, singrisone@gmail.com
Jewsbury, Paul Adrian, Educational Testing Service, pjewsbury@ets.org
Ji, Feng, fengji@berkeley.edu
Jia, Yue, Educational Testing Service, yjia@ets.org
Jiang, Yang, Educational Testing Service, yj2211@tc.columbia.edu
Jiao, Hong, University of Maryland, hjiao@umd.edu
Jin, Kuan-Yu, Hong Kong Examinations and Assessment Authority, kyjin@hkeaa.edu.hk
Jin, Rong, Riverside Insights, rongjin2012@gmail.com
Jin, Ying, Association of American Medical Colleges, yjin@aamc.org
Jing, Shumin, shumin-jing@uiowa.edu
Jodoin, Michael, National Board of Medical Examiners, mjodoin@nbme.org
Johnson, David, University of Minnesota, johns006@umn.edu
Johnson, Matthew, ETS, msjohnson@ets.org
Jones, Andrew, American Board of Surgery, ajones@absurgery.org
Jones, Peggy, Pasco County (FL) Dist School, pejones@pasco.k12.fl.us
Joo, Seang-Hwane, Educational Testing Service, sjoo001@ets.org

Ju, Unhee, Riverside Insights, early0522@gmail.com
Jung, Hyun Joo, hyunjoo.jung2@gmail.com
Justus, Brandon Johnathan, brandon.justus@kpu.ca
Kachchaf, Rachel R, rachel.kachchaf@smarterbalanced.org
Kaira, Leah, Pearson, leahkaira@gmail.com
Kaliski, Pamela, ABIM, PKaliski@ABIM.ORG
Kamata, Akihito, akamata@gmail.com
Kane, Joanne, National Conference of Bar Examiners, jkane@ncbex.org
Kane, Michael, ETS
Kang, Hyeon-Ah, University of Texas at Austin, hkang@austin.utexas.edu
Kanopka, Klint, kkanopka@stanford.edu
Kao, Shu-chuan, NCSBN, skao@ncsbn.org
Karamese, Hacer, hacer-karamese@uiowa.edu
Kartal, Gamze, University of Illinois at Urbana-Champaign, gkartal2@illinois.edu
Kasli, Murat, University of Miami, muratkasli@miami.edu
Katz, Daniel, University of California, Santa Barbara, dkatz@education.ucsb.edu
Kehinde, Olasunkanmi, Washington State University, kehinde.james@wsu.edu
Kelberlau, Darin, Millard Public Schools, dckelberlau@mpsomaha.org
Keng, Leslie, Center for Assessment, lesliekeng@gmail.com
Kern, Justin L., University of Illinois at Urbana-Champaign, kern4@illinois.edu
Ketterlin Geller, Leanne, Southern Methodist University, lkgeller@mail.smu.edu
Keum, Eunhee, UCLA CRESST, keum@cresst.org
Kim, Eunbee, Georgia Institute of Technology, eunbee.kim@gatech.edu
Kim, Hyung Jin, University of Iowa, hyungjin-kim@uiowa.edu
Kim, Jungnam, NWEA, jungnam95@hotmail.com
Kim, Seongeun, University of North Carolina, waytogokim@gmail.com
Kim, Seungman, seungman.kim@ttu.edu
Kim, Sooyeon, ETS, skim@ets.org
Kim, Stella, University of North Carolina at Charlotte, stella-kim@uncc.edu
Kim, Young Yee, American Institutes for Research, ykim@air.org
King, Jacqueline, James Madison University, king5je@dukes.jmu.edu
Kingsbury, G. Gage, gagekingsbury@comcast.net
Kingston, Neal, University of Kansas, nkingsto@ku.edu
Klugman, Emma M., Harvard Graduate School of Education, eklugman@g.harvard.edu
Knezevich, Lily, Law School Admission Council, lknezevich@lsac.org
Knight, Brooke, Scintilla Charter Academy, bknight@scintillacharteracademy.com
Knight, Decca, decca.knight@me.com
Kolen, Michael, University of Iowa, kolenmichael@gmail.com
Kollias, Charalambos, National Foundation for Educational Research, chkollias@outlook.com
Konold, Tim, University of Virginia, konold@virginia.edu
Koretz, Daniel, Harvard Graduate School of Education, daniel_koretz@harvard.edu
Kosh, Audra, Edmentum, audrakosh@gmail.com
Kraus, Ken, National Conference of Bar Examiners, kkraus@ncbex.org

Kroopnick, Marc, Association of American Medical Colleges, mkroopnick@aamc.org
Krost, Kevin, Fralin Biomedical Research Institute, kevinkrost@vt.edu
Kuang, Huan, University of Florida, huan2015@ufl.edu
Kukea Schultz, Pohai, University of Hawaii, pohai@hawaii.edu
Kuklick, Livia, IPN Kiel, kuklick@leibniz-ipn.de
Kuo, Tzu-Chun, Cambium Assessment, Inc., june.kuo@cambiumassessment.com
Kwon, Tae Yeon, taeyeon.kwon@ufl.edu
Kyllonen, Patrick Charles, ETS, pkyllonen@ets.org
LaFlair, Geoff, Duolingo, geoff@duolingo.com
Lambert, Richard, UNC Charlotte, rglamber@uncc.edu
Lane, Suzanne, University of Pittsburgh, sl@pitt.edu
Langenfeld, Thomas E., TEL Measurement, telangenfeld@gmail.com
Lau, Sok-Han, University of Hawaii at Manoa, sokhan@hawaii.edu
Lay, Alexandra, almart25@uncg.edu
Lee, Bitna, Kyungpook National University, lllj226@naver.com
Lee, Bob, Massachusetts Department of Education, Bob.Lee@mass.gov
Lee, Chansoon, Liberty University, clee180@liberty.edu
Lee, Dukjae, University of Massachusetts Amherst, dlee@umass.edu
Lee, Haeju, HLEE@uncg.edu
Lee, Juyeon, jl13335@uga.edu
Lee, Mina, University of Massachusetts Amherst, mina.mh.lee@gmail.com
Lee, Soo, American Institutes for Research, slee@air.org
Lehman, Blair, Educational Testing Service, blehman@ets.org
Lehrfeld, Jon, Educational Testing Service, jlehrfeld@ets.org
Leventhal, Brian C, James Madison University, leventbc@jmu.edu
Lewis, Daniel, Creative Measurement Solutions LLC, dan.lewis@creativemeasurement.com
Lewis, Jennifer L., University of Massachusetts Amherst, jlewi0@umass.edu
Li, Dongmei, ACT, dongmei.li@act.org
Li, JingYi, Beijing Normal University, 201921630013@mail.bnu.edu.cn
Li, Jun, University of Minnesota Twin Cities, lixx1474@umn.edu
Li, Lanrong, jessicalilr2011@gmail.com
LI, Xinyue, Penn State University, xql5285@psu.edu
Li, Zhushan Mandy, Boston College, zhushan.li@bc.edu
Liang, Qianru, University of Hong Kong, liangqr@hku.hk
Liao, Dandan, Cambium Assessment, Inc., dandan.liao@cambiumassessment.com
Liao, Manqian, mancyliao@gmail.com
Lim, Hwanggyu, Graduate Management Admission Council, hglim83@gmail.com
Lin, Qiao, University of Illinois at Chicago, qlin7@uic.edu
Lin, Ye, Ascend Learning, yelin.nora@gmail.com
Lin, Zhongtian, Cambium Assessment, Inc, zhongtian.lin@cambiumassessment.com
Lindner, Marlit Annalena, IPN Kiel, mlindner@ipn.uni-kiel.de
Liu, Chunyan, National Board of Medical Examiners, cliu@nbme.org
Liu, Na, na.liu@gatech.edu

Liu, Ou Lydia, ETS, lliu@ets.org
Liu, Xiang, Educational Testing Service, xliu003@ets.org
Liu, Xiaowen, xiaowen.liu@uconn.edu
Liu, Yang, University of Maryland, College Park, yliu87@umd.edu
Liu, Yu, yliu107@uh.edu
Lorie, William A., Center for Assessment, william.lorie@gmail.com
Lottridge, Susan, Cambium Assessment, susanlottridge@hotmail.com
Lu, Chang, University of Alberta, clu4@ualberta.ca
Lu, Jing, Northeast Normal University, luj282@nenu.edu.cn
Lu, Ru, Educational Testing Service, rlu@ets.org
Lu, Yi, Federation of State Boards of Physical Therapy, ylu@fsbpt.org
Luecht, Richard Melvin, UNC Greensboro, rmluecht@uncg.edu
Luo, Yong, jackyluoyong@gmail.com
Lyons, Susan, Women in Measurement, Inc., susan@womeninmeasurement.org
Ma, Mingjia, University of Iowa, mingjia-ma@uiowa.edu
Ma, Ye, University of Iowa, ye-ma@uiowa.edu
MacInnes, Joshua, Ascend Learning, macjosh1122@gmail.com
Madison, Matthew James, University of Georgia, mjmadison@uga.edu
Malatesta, Jaime, Graduate Management Admission Council, jmalatesta@gmac.com
Man, Kaiwen, University of Alabama, kman@ua.edu
Mardones, Constanza, University of Georgia, cam04214@uga.edu
Marion, Scott, Center for Assessment, smarion@nciea.org
Maris, Gunter, ACT, Gunter.Maris@act.org
Martinez, Jose Felipe, UCLA - School of Education and Information Studies, jfmtz@ucla.edu
Matthew, Nirupa, Curriculum Associates, nmatthew@cainc.com
Mayfield, Kerrita, Amherst Public Schools, kteachergirl@hotmail.com
McCaffrey, Daniel, Educational Testing Service, dmccaffrey@ets.org
McCall, Martha, McKinsey & Company, mccall.marty@gmail.com
McCallister, Cynthia, New York University, mccallistercynthia@gmail.com
McClarty, Katie, Renaissance, Katie.McClarty@renaissance.com
McCormick, Carina M., Buros Center for Testing, cmccormick@buros.org
McNamara, Danielle, Arizona State University, dsmcnamara1@gmail.com
Mee, Janet, NBME, jmee@nbme.org
Meng, Huijuan, AWS, huijuan_meng@hotmail.com
Meyer, Patrick, NWEA, meyerjp3@gmail.com
Miao, Jing, Educational Testing Service, jmiao@ets.org
Michel, Rochelle, Curriculum Associates, rochelle.michel@gmail.com
Middlestead, Andrew J., Michigan Department of Education, middlesteada@michigan.gov
Middleton, Kyndra, Howard University, kvmiddleton@gmail.com
Miller, Amanda, Scintilla Charter Academy, amiller@scintillacharteracademy.com
Mills, Christine, Ascend Learning, Christine.Mills@ascendlearning.com
Mills, Craig, NBME, cmills@nbme.org
Mintz, Catherine Elizabeth, University of Iowa, catherinemathers93@gmail.com

Mittapalli, Kavita, MN Associates, Inc., mailto:kmittapalli@gmail.com
Molenaar, Dylan, University of Amsterdam, d.molenaar@uva.nl
Moncaleano, Sebastian, Boston College, moncaleano91@gmail.com
Monroe, Scott, University of Massachusetts Amherst, smonroe@educ.umass.edu
Monteiro, Elissa Mara, University of California, Riverside, emont062@ucr.edu
Montgomery, Melinda, NWEA, melindasmontgomery@gmail.com
Moore, Joann, ACT, joannlmoore@gmail.com
Morell, Linda, lindamorell@berkeley.edu
Morell, Monica, University of Maryland, mmorell@umd.edu
Morin, Maxim, maxim.morin.13@gmail.com
Morrison, Kristin M., Curriculum Associates, KMorrison@cainc.com
Moses, Tim, College Board, tmoses@collegeboard.org
Moyer, Eric, Pearson, eric.moyer@utexas.edu
Muckle, Timothy, Board of Pharmacy Specialties, tmuckle@aphanet.org
Muntean, William J, National Council of State Boards of Nursing, williamjmuntean@gmail.com
Myers, Aaron, University of Arkansas, ajm045@uark.edu
Nash, Brooke, University of Kansas, bnash@ku.edu
Naveiras, Matthew David, Peabody College of Vanderbilt, matthew.d.naveiras@vanderbilt.edu
Nemeth, Yvette, HumRRO, ynemeth@humrro.org
Nichols, Paul, NWEA, paul.nichols@nwea.org
Nisbet, Isabel, University of Cambridge, nisbet.isabel@gmail.com
Niu, Chunling Chunling, University of Kentucky, chunling.niu@gmail.com
O'Donnell, Francis, National Board of Medical Examiners, fodonnell@nbme.org
Ogut, Burhan, American Institutes for Research, bogut@air.org
Olgar, Suleyman, Florida Department of Education, suleymanolgar@yahoo.com
Oliveri, Maria Elena, Buros Center for Testing-UNL, oliveri.m@live.com
Ong, Thai Quang, National Board of Medical Examiners, tong@nbme.org
O'Riordan, Maura, University of Massachusetts Amherst, moriordan@umass.edu
Orona, Gabe Avakian, University of California, Irvine, gorona@uci.edu
Padgett, Robert N, noah_padgett1@baylor.edu
Paek, Insu, Florida State University, ipaek@fsu.edu
Paek, Pamela, pamelapaek@gmail.com
Pan, Qianqian, University of Hong Kong, panqianqian2013@gmail.com
Pan, Yiqin, University of Wisconsin-Madison, pan74@wisc.edu
Pandian, Ravi, National Board of Medical Examiners, rpandian357@gmail.com
Park, Seohee, hee6904@gmail.com
Park, Sook Hyun, spark01@utexas.edu
Parker, John, Floyd County Schools, jparker@floydboe.net
Pastor, Dena, James Madison University, pastorda@jmu.edu
Patarapichayatham, Chalie, Southern Methodist University, cpatarapichy@smu.edu
Patelis, Thanos, Fordham University, Teachers College, University of Kansas, tpatelis@yahoo.com
Patton, Elizabeth Adele, BPatton@cainc.com
Patz, Richard, University of California, Berkeley, rpatz@berkeley.edu

Paulsen, Justin, HumRRO, JPaulsen@humrro.org
Peabody, Michael R, National Association of Boards of Pharmacy, michael.peabody77@gmail.com
Peasley, Donald, U.S. Department of Education, Donald.Peasley@ed.gov
Pellegrino, James, University of Illinois at Chicago, pellegjw@uic.edu
Peng, Fang, National Council of State Boards of Nursing, pfrenee@gmail.com
Perie, Marianne, Measurement in Practice, LLC, mp@measurementinpractice.com
Perkins, Beth, James Madison University, perkinba@jmu.edu
Pham, Duy N., Educational Testing Service, dnpham@ets.org
Picou, Aigner, The Learning Agency, aigner@the-learning-agency.com
Pierre-Louis, Medjy, mepi9219@colorado.edu
Poe, Mya, Northeastern University, m.poe@northeastern.edu
Pointner, Julie, DRC, juliekorts3@gmail.com
Porter, Andrew, University of Pennsylvania, andyp@gse.upenn.edu
Prier, Darius, Duquesne University, prierd@duq.edu
Qiao, Xin, xin.qiao56@gmail.com
Qin, Qi, Gwinnett County Public Schools, qinqi715@gmail.com
Quan, Jia, jia.quan@ufl.edu
Rabinowitz, Stanley N, Pearson, stanley.rabinowitz@pearson.com
Raczynski, Kevin, University of Georgia, kevin.raczynski@gmail.com
Raddatz, Mikaela, American Board of Physical Medicine and Rehabilitation, MRaddatz@abpmr.org
Randall, Jennifer, University of Massachusetts, jrandall@educ.umass.edu
Rasooli, Amir, amir.rasooli@queensu.ca
Rebouças-Ju, Daniella, University of Notre Dame, drebouca@nd.edu
Reckase, Mark, Psychometric Solutions, reckase@msu.edu
Reid, Aileen, UNC Greensboro, amreid3@uncg.edu
Reshetar, Rosemary, National Conference of Bar Examiners, rreshetar@ncbex.org
Rewley, Kelly, American Board of Internal Medicine, KRewley@abim.org
Rios, Joseph A., University of Minnesota, jrios@umn.edu
Rivera, Christopher, East Carolina University, RIVERAC@ECU.EDU
Roeber, Edward Dean, Michigan Assessment Consortium, roeber@msu.edu
Rollins, Jonathan, West Virginia Department of Education, rollinsj14@gmail.com
Rome, Logan, Curriculum Associates, lrome@cainc.com
Rosales De Veliz, Leslie Vanessa, Juarez & Associates, lrosales@juarezassociates.com
Rosenberg, Sharyn, NAGB, sharyn.rosenberg@ed.gov
Rubright, Jonathan, National Board of Medical Examiners, jrubright@nbme.org
Runyon, Christopher, NBME, CRunyon@nbme.org
Rupp, Andre, Mindful Measurement, dr.andre.rupp@gmail.com
Rutkowski, Leslie, Indiana University, lrutkows@iu.edu
Sahin, Fusun, American Institutes for Research, fsahin@air.org
Sahin, Sakine Gocer, WIDA at UW-Madison, gocersahin@wisc.edu
Sarac, Merve, University of Wisconsin-Madison, sarac@wisc.edu
Satkus, Paulius, James Madison University, satkuspx@jmu.edu
Sato, Edynn, Sato Education Consulting LLC, edynn@satoeducationconsulting.com

Sauder, Derek, American Board of Internal Medicine, dsauder@abim.org
Scalise, Kathleen, University of Oregon, kscalise@uoregon.edu
Schellman, Madeline, mas13@uga.edu
Schlax, Jasmin, Johannes Gutenberg University, jaschlax@uni-mainz.de
Schnabel, Sarah, American Board of Ophthalmology, Sarah.d.schnabel@gmail.com
Schneider, Christina, NWEA, christina.schneider@nwea.org
Schneider, Wei, shuang-wei@uiowa.edu
Sgammato, Adrienne, ETS, asgammato@ets.org
Shaw, Emily, College Board, eshaw@collegeboard.org
Shaw, Robert C., National Board for Respiratory Care, robert.shaw@nbrc.org
Shear, Benjamin R., University of Colorado Boulder, benjamin.shear@colorado.edu
Shen, Yawei, ys37335@uga.edu
Shepard, Lorrie Ann, University of Colorado Boulder, Lorrie.Shepard@Colorado.edu
Shermis, Mark David, American University of Bahrain, mshermis@gmail.com
Shi, Qingzhou, qshi7@crimson.ua.edu
Shin, Hyo Jeong, Educational Testing Service, hshin@ets.org
Shin, Jinnie, jinnie.shin@coe.ufl.edu
Shin, Nami, UCLA CRESST, nami0623@gmail.com
Shin, Sujie, California Collaborative for Educational Excellence, sshin@ccee-ca.org
Sikali, Emmanuel, Emmanuel.Sikali@ed.gov
Silva, Monica, Pontificia Universidad Católica de Chile, msilvara@uc.cl
Sinclair, Andrea, HumRRO, asinclair@humrro.org
Sinharay, Sandip, Educational Testing Service, ssinharay@ets.org
Sipahi, Rabia Esma, University of Kansas, rabiasipahi@ku.edu
Sireci, Stephen, University of Massachusetts Amherst, sireci@acad.umass.edu
Skorupski, Billy, Questar Assessment, wskorupski@questarai.com
Smith, Jeffrey, Township High School District 214, jeffrey.smith@d214.org
Smith, Jessalyn, DRC, jsmith@datarecognitioncorp.com
Smith, Michael Joseph, University of Virginia, mjs9t@virginia.edu
Smith, Mireya Carmen-Martinez, University of Minnesota, mart1799@umn.edu
Soland, James, University of Virginia, jgs8e@virginia.edu
Somay, Su, NBME, SSomay@nbme.org
Song, Yoon Ah, Center for Applied Linguistics, episteme84@hotmail.com
Sparks, Anthony, asparks@smu.edu
Spitz, Deborah, U.S. Department of Education, Deborah.Spitz@ed.gov
Stanley, Arthur, Pearson, arthur.stanley@pearson.com
Steele, Taisha, Pearson, Tasha.steele@pearson.com
Steedle, Jeffrey, ACT, jtsteedle@gmail.com
Stickney, Eric, Renaissance, Eric.Stickney@renaissance.com
Stiggins, Richard, Assessment Training Institute, ricks@assessmentinst.com
Stone, Alexandra, University of Connecticut, alexandra.stone@uconn.edu
Student, Sanford, University of Colorado Boulder, sanford.student@colorado.edu
Su, Hong, China National Institute of Education Sciences, suh@nies.net.cn

Su, Kun, UNC Greensboro, kun.su518@gmail.com
Suksiri, Weeraphat, University of California, Berkeley, w.suksiri@berkeley.edu
Sun, Huaping, American Board of Anesthesiology, sonkahe@hotmail.com
Sun, Yan, Rutgers University, yan.sun@rutgers.edu
Svetina Valdivia, Dubravka, Indiana University, dsvetina@indiana.edu
Tang, Nai-En, naientang@gmail.com
Tang, Steven, eMetric LLC, steven@emetric.net
Tang, Xiuxiu, tang469@purdue.edu
Tang, Xueying, University of Arizona, xytang@math.arizona.edu
Taylor, Melinda Ann, ACT, taylor.melinda@gmail.com
Thacker, Arthur, HumRRO, athacker@humrro.org
Thompson, Nathan, Assessment Systems Corporation, nthompson@assess.com
Thompson, W. Jake, University of Kansas, wjakethompson@gmail.com
Thum, Yeow Meng, NWEA, yeow.meng@nwea.org
Thurlow, Martha, National Center on Educational Outcomes, thurl001@umn.edu
Tian, Chen, 736218349@qq.com
Tong, Ye, Pearson, ye.tong@pearson.com
Topczewski, Anna, WestEd, topczewski.anna@gmail.com
Torres Irribarra, David, Pontificia Universidad Católica de Chile, davidtorres@uc.cl
Toton, Sarah Linnea, Caveon Test Security, sarah.toton@caveon.com
Traynor, Anne, Purdue University, atraynor@purdue.edu
True, Rhonda, Nebraska Department of Education, Rhonda.True@nebraska.gov
Twing, Jon S., Pearson, jon.s.twing@pearson.com
Ulitzsch, Esther, Leibniz Institute for Science and Mathematics Education, ulitzsch@inp.uni-kiel.de
Underwood, Sarah, Florida Department of Education, Sarah.Underwood@fldoe.org
van Bork, Riet, Center for Philosophy of Science, University of Pittsburgh, rietvanbork@hotmail.com
Vance, Amelia, Future of Privacy Forum, avance@fpf.org
van Rijn, Peter, ETS Global, pvanrijn@etsglobal.org
Vassileva, Victoria, Arthur AI, victoria@arthur.ai
Verges, Vince, Florida Department of Education, vergesvincent@gmail.com
Villafuerte, Catherina, University of Connecticut, catherina.villafuerte@uconn.edu
Vo, Yen, University of Iowa, yen-vo@uiowa.edu
von Davier, Alina A, Duolingo, avondavier@duolingo.com
von Davier, Matthias, Boston College, vondavim@bc.edu
Walker, Cindy M, Research Analytics Consulting LLC, dr.cindy.m.walker@gmail.com
Walker, Michael E., Educational Testing Service, mwalker@ets.org
Wall, Nathan, eMetric, nwall@emetric.net
Wan, Lei, wanlei2254@yahoo.com
Wan, Siyu, University of Massachusetts Amherst, siyuwan@umass.edu
Wang, Aijun, FSBPT, wajlm2003@gmail.com
Wang, Chun, University of Washington, wang4066@uw.edu
Wang, Jiaxi, Rutgers University, jw1218@scarletmail.rutgers.edu
Wang, Lu, ACT, lu.wang@act.org

Wang, Nan, nw13c@my.fsu.edu
Wang, Nixi, nixiwang@uw.edu
Wang, Shichao, ACT, shichao.wang@act.org
Wang, Shiyu, University of Georgia, swang44@uga.edu
Wang, Shudong, NWEA, shudong.wang@NWEA.org
Wang, Songtao, University of Victoria, songtaowang@uvic.ca
Wang, Ting, American Board of Family Medicine, twang@theabfm.org
Wang, Weimeng, University of Maryland, College Park, wwang111@umd.edu
Wang, Yang Caroline, Education Analytics, cwang@edanalytics.org
Wang, Yibo, yibo-wang@uiowa.edu
Wang, Yu, Rutgers University, yw741@scarletmail.rutgers.edu
Wang, Zhaoyu, Georgia Institute of Technology, zwang3036@gatech.edu
Wang, Zhen, Cambium Assessment, zhen.wang@cambiumassessment.com
Washington, Ernest, University of Massachusetts Amherst, ewashington@educ.umass.edu
Weeks, Jonathan, Educational Testing Service, jweeks@ets.org
Wei, Hsin-Ro, Riverside Insights, Hsin-ro.wei@riversideinsights.com
Weiner, John, PSI Services, LLC, jweiner@psionline.com
Weir, J. B., National Commission on Certification of Physician Assistants, weirjb@gmail.com
Weissman, Alexander, Law School Admission Council, aweissman@lsac.org
Wells, Amy Stuart, Teachers College, wells@exchange.tc.columbia.edu
Wheeler, Jordan M., University of Georgia, jmwheeler@uga.edu
Wheeler, Kelley, ACS Ventures, LLC, kelleyrwheeler@gmail.com
White, Jennifer, Floyd County Schools, jwhite@floydboe.net
White, Lauren, Florida Department of Education, Lauren.White@fldoe.org
Whitmer, John, Federation of American Scientists, jwhitmer@fas.org
Wiberg, Marie, Department of Statistics, USBE, marie.wiberg@umu.se
Wiley, Andrew, ACS Ventures, LLC, Awiley999@gmail.com
Wiley, Caroline, Educational Testing Service, ecwylie@ets.org
Williams, Ashley, Bioplicity, ash.blake.williams@gmail.com
Williamson, David, College Board, dwilliamson215a@gmail.com
Wilson, Mark, University of California, Berkeley, markw@berkeley.edu
Wind, Stefanie A., University of Alabama, swind@ua.edu
Winter, Phoebe C, phoebe.winter@outlook.com
Wise, Steven, NWEA, steve.wise@nwea.org
Wollack, James, University of Wisconsin, jwollack@wisc.edu
Woo, Ada, Ascend Learning, adawoo811@gmail.com
Word, Andrea, University of Alabama in Huntsville, worda@uah.edu
Workman, Trent, Pearson, Trent.Workman@Pearson.com
Wu, Tong, twu11@uncc.edu
Wu, Tong, wu473@purdue.edu
Wu, Yi-Fang, ACT, Yi-Fang.Wu@act.org
Wyse, Adam, Renaissance, adam.wyse@renaissance.com
Xi, Nuo, VIPKID, nuoxi.1@gmail.com

Xiong, Jiawei, University of Georgia, jx56584@uga.edu

Xu, Guanlan, guanlan-xu@uiowa.edu

Xu, Jiajun, University of Georgia, jiajunxu@uga.edu

Xu, Shuangshuang, renixu@umd.edu

Xu, Wei, National Council of State Boards of Nursing, x.wei1007@gmail.com

Yan, Yan, Georgia Tech, yany@gatech.edu

Yancey, Kevin, Duolingo, kyancey@duolingo.com

Yaneva, Victoria, National Board of Medical Examiners, vyaneva@nbme.org

Yang, Ji Seung, University of Maryland, jsyang@umd.edu

Yang, Yi, Columbia University, yi.y@columbia.edu

Yavuz, Sinan, University of Wisconsin-Madison, syavuz@wisc.edu

Ye, Tianpeng, tianpeng-ye@uiowa.edu

Yildirim-Erbasli, Seyma Nur, University of Alberta, seymanur@ualberta.ca

Yu, Nan Sook, Chonnam National University, nansooksb@gmail.com

Yuan, Haimiao, University of Iowa, heemeol@gmail.com

Yuan, Kun, Association of American Medical Colleges, kyuan@aamc.org

Zehner, Fabian, DIPF | Leibniz Institute for Research and Information in Education, Centre f. Int. Student Assessm., fabian.zehner@dipf.de

Zeng, Biao, Beijing Normal University, biaozeng@mail.bnu.edu.cn

Zeng, Ji, Michigan Department of Education, zengj@michigan.gov

Zeng, Wen, Cambium Assessment, Inc., wen.zeng@cambiumassessment.com

Zeng, Yifang, yifang.zeng@ttu.edu

Zenisky, April, University of Massachusetts Amherst, azenisky@educ.umass.edu

Zepeda, Sandra Cecilia, Universidad Catolica, sandrazepeda@gmail.com

Zhan, Peida, pdzhan@gmail.com

Zhang, Jihong, University of Iowa, jihong-zhang@uiowa.edu

Zhang, Liru, Assessment Consulting Services, liru.zhang@outlook.com

Zhang, Mingqin, University of Iowa, mingqin-zhang@uiowa.edu

Zhang, Mo, Educational Testing Service, mzhang@ets.org

Zhang, Susu, University of Illinois at Urbana-Champaign

Zhang, Ting, AIR, tzhang@air.org

Zheng, Xiaying, American Institutes for Research, xzheng@air.org

Zhu, Mengxiao, Educational Testing Service, mzhu@ets.org

Zhu, Shuai, TAL Education Group, 2536751571@qq.com

Zhu, Zhemin, Beihua University, zhemin.zhu@foxmail.com

Zopluoglu, Cengiz, University of Oregon, cengiz@uoregon.edu

Zor, Selay, University of Georgia, sz37952@uga.edu

Zurkowski, Joyce, zurkowski_j@cde.state.co.us

Zwick, Rebecca, Educational Testing Service, RZwick@ETS.ORG

Schedule At a Glance

Pre-conference Sessions

	Tuesday May 18, 2021		**Thursday May 20, 2021**	
11:00a - 12:30p	A Case Study in Measurement Practice and The Public Perception	(Invited Session) Lessons about the modeling and measurement of human abilities	Unpacking Cognitive Complexity: What is it and Why is it so Hard?	(Invited Session) Using Longitudinal Assessment to Support Professional Development
2:00p - 3:30p	Embedded Standard Setting: Research & Advances	Psychometric Challenges and Potential Solutions for Educator Testing in Pandemic Environment	Procedures for establishing and evaluating linkages between scores collected in different modes	Ethics & STEM Assessments: Content modeling, construct mapping, psychometric models, mitigating bias
	Tuesday May 25, 2021		**Thursday May 27, 2021**	
11:00a - 12:30p	(SIGIMIE Session) Current Challenges in Large-scale Assessment and Responses/Innovations	Fair and Valid Assessment of ELs with the Most Significant Cognitive Disabilities	Psychometrics for Digital-First Assessments: The Duolingo English Test Application	Assessing COVID-19 Impacts on Assessment and Learning using Star Interim Assessments
2:00p - 3:30p	(SIGIMIE Session) Scaling, Linking, & Equating Du Jour: A Discussion with Experts	(Invited Session) Advancing Women in Measurement: Barriers and Opportunities	Large-scale Educational Data Sets and the Ethics of their Monetization	(Invited Session) Education literacy for psychometricians
	Tuesday June 1, 2021		**Thursday June 3, 2021**	
11:00a - 12:30p	On the Assessment of Non-Cognitive Competencies in Licensure: Why, Whether, and How?	(SIGIMIE Session) Building a Multidimensional Future: A Conversation on Big Data and Educational Measurement	(SIGIMIE Session) Testing Time: The Push and Pull in High-Stakes State Accountability Assessments	Advancing Assessment in Medical Education
2:00p - 3:30p	(SIGIMIE Session) Debating the training of future measurement professionals	Guidelines for Technology-Based Assessments: An ITC and ATP Collaboration	(SIGIMIE SESSION) Challenges and opportunities in delivering virtual oral and OSCE examinations	Creating Coherence: Integrating Principled Assessment Design, PLDs, and Standard Setting

Conference Week: Wednesday June 9

9:00a - 10:30a	(Invited Session) Using Educational Assessments to Educate: Opportunities for Leveraging the "Power" of Assessment	From CAT to Smart Learning – Urgent Research During the Pandemic	(Invited Session) A.I. and Machine Learning	High Definition Detection of Testing Misconduct	Issues in Item and Test Design	The Resurgence of Interim Assessment—Bringing Teaching and Testing Back Together	Innovations in Response Time Models
10:35a - 11:00a	Coffee Chat Session #1	Coffee Chat Session #2	Coffee Chat Session #3	Coffee Chat Session #4	Coffee Chat Session #5	Coffee Chat Session #6	Coffee Chat Session #7
11:15a - 12:45p	(CODIT Feature Session) Black Lives Matter in Educational Measurement	(SIGIMIE Session) Navvy Education: Building and Implementing a Statewide Diagnostic Assessment System	Advancing Digital Instruction and Assessment with Natural Language Processing & Learning Analytics	Psychometric Modeling of Data Based on a Table of Specifications	Validity, Psychometric Properties, and Accessibility of Innovative Item Types in K-12 Assessments	Application of Fit Statistics	Application of Response Time Models
1:00p - 2:00p	Electronic Board Session #1	(Research Blitz) Focus on Linking and Equating	(Research Blitz) Focus on Adaptive Testing	Item Evaluation Strategies	(Invited Session) The value of assessment data from spring 2021: A debate	Focus on Students with Disabilities	Automatic Item Generation Considerations
2:15p - 3:45p	Modeling Response Time: A Collaborative Case Study on a High-Stakes Admission Exam	Developing Successful and Impactful Assessment Products – Balancing Research and Business Considerations (Joint Session with Association of Test Publishers)	(Invited Session) Assessment Literacy: Practical Applications and Implications (National Association of Assessment Directors)	Going beyond Scores: Understanding Response and Process in Large-scale Assessments	Fostering Assessment Quality: Learning from Federal "Peer Review" Criteria, Process, and Impact	Topics in Standard Setting	Grading and Raters
4:00p - 5:30p	Standard Setting Challenges and Solutions for Innovative Assessment System Designs	Scrutinizing item responses and response times: Experimental and analytic approaches	Suggestions for Fairness and Equity, as well as Quality, in Testing	Electronic Board Session #4	Diagnostic Assessments: Moving from Theory to Practice	Topics in Measuring Growth	Techniques in Machine Learning or Artificial Intelligence

Conference Week: Thursday June 10

9:00a - 10:30a	Focus on English Language Learners	The Past, Present, and Future of Item Difficulty Modeling	Leveraging Response Process Data to Support Testing Programs: Strategies and Real-world Examples	Using Artificial Intelligence for Constructed-Response Scoring: Some Practical Considerations	Topics in Item Response Theory	The Future of K-12 Assessment: Is there One?	Applications of Diagnostic Classification Models
10:45a - 12:45p	NCME Business Meeting						
1:00p - 2:00p	(Invited Session) Stakeholder Perspectives on Validating Licensure Examinations	(Research Blitz) Topics in Test Development	(Research Blitz) Techniques for Missing Data and Guessing Behavior	Topics in Multidimensional Item Response Theory	Involve me and I learn: Applying culturally responsive assessment practices to equitably measure learning of Indigenous students in North America	Graduate Student Electronic Board Session	Generalizability Theory Applications
2:15p - 3:45p	Leveraging Process Information in International Large-Scale Assessments: Recent Findings from PIAAC	(Invited Session) Assessments For Different Purposes: Issues on Scoring, Score Use, and Measurement	The Impact of COVID-19 on Educational Measurement, Part 1: K-12 Assessment	Developing an Alternate English Language Proficiency Assessment within a Principled Design Framework	Topics in Validity	Differential Item Functioning (DIF) Applications	Applications in Adaptive Testing
4:00p - 5:30p	(Invited Session) Pivoting in a Pandemic	Probabilistic Graphical Models for Writing Process Data	Mode Comparability in College Admissions Testing: In-depth Investigations and Methodological Considerations	An Assessment Development and Management (ADM) System for Educational Applications.	Multistage Testing with Multiple Subscales: An Investigation of Design and Analysis	Impact of Test Design and Features on Performance	Security Issues in Credentialing

Conference Week: Friday June 11

9:00a - 10:30a	Remembering "Career" in College and Career Readiness	Diving into NAEP Process Data to Understand Students' Test Taking Behaviors	(Invited Session) Lessons Learned from the Pandemic: How do credentialing programs prepare for the next major crisis/disruption?	Practical Issues in Automated Test Assembly	Recent Research on Detecting Disengaged Test Taking	Communicating results	The Impact of COVID-19 on Educational Measurement, Part 2: Admissions and Certification
10:35a - 11:00a	Coffee Chat Session #1	Coffee Chat Session #2	Coffee Chat Session #3	Coffee Chat Session #4	Coffee Chat Session #5	Coffee Chat Session #6	Coffee Chat Session #7
11:15a - 12:45p	Developing a Longitudinal Assessment: Using Innovations and Research to Address Measurement Issues	Electronic Board Session #3	Identifying Rushing in CAT and Investigating the Effects on Differentiated Instruction	Designing and Evaluating Innovative Assessment Systems: Combining Research and Practice	The Value of and Values in Educational Assessment	PISA and TIMSS Topics	Adaptive Testing Topics
1:00p - 2:00p	(Invited Session) Looking ahead – Bridging future research and practice in credentialing	(Research Blitz) Focus on CDM and DCM	(Research Blitz) Item Response Theory Applications	(Research Blitz) Measurement of Transacademic Skills	(Invited Session) Where Do We Go from Here? A Practitioner's Discussion of Our Post-Pandemic World	Electronic Board Session #2	Going for Broke: Acknowledging and Disrupting the Barriers to Black Lives Mattering in Measurement
2:15p - 3:45p	(Invited Session) The Future of College Admissions Testing	Rosetta Stone or Tower of Babel? Debating methods for NAEP-linked aggregate scores	Impact of COVID-19 on Assessment	Applications of Process Data	The AERA/APA/NCME Standards: Is it time to revisit the policy of self-enforcement?	Fairness for All Examinees	Security issues in Education
4:00p - 5:30p	Exploring Scoring and Psychometric Modeling Strategies for Innovative Items	IRTree Models: The Illus-tree-ous and In-tree-guing Response Process Models	Innovations in Detection of Test Collusion	Moving Bookmark Standards Setting from In-Person to Virtual: Best Practices/Lessons Learned	$E \neq MC^2$: Equity doesn't equal measurement calibration squared	Topics in Linking and Equating	Topics in Cognitive Diagnosis Modeling

Gold Level

Silver Level

Friend Level

Save the Date for NCME 2022!
April 22 – April 25
San Diego, CA

www.ncme.org

OUR MISSION

The National Council on Measurement in Education is a community of measurement scientists and practitioners who work together to advance theory and applications of educational measurement to benefit society.