NCME national council on measurement in education

2018 Annual Meeting
April 12-16
New York, NY

Here and There and Back Again: Making Assessment a Stronger Force for Positive Impact on Teaching and Learning

NCME national council on measurement in education

National Council on Measurement in Education

Here and There and Back Again:
Making Assessment a Stronger Force for
Positive Impact on Teaching and Learning

2018 Training Sessions
April 12-13

2018 Annual Meeting
April 14-16

Westin New York at Times Square New York, NY

#NCME18

Welcome from the Program Chairs

Welcome to New York, welcome to New York! Welcome, friends and colleagues, to the 2018 NCME Annual Meeting. We are pleased to present you with this year's NCME program. Our goal in putting together this slate of sessions has been balance: we have sought to represent research from different testing contexts, from a wide range of perspectives, from behind-the-scenes test development efforts across topics in our field to activities that advance the ways that tests and test results can be made accessible and useful to stakeholders.

This year's conference theme of "Here and There and Back Again: Making Assessment a Stronger Force for Positive Impact on Teaching and Learning" seeks to cultivate the interplay between testing (in all its forms) and the processes of instruction and learning. Carrying on with NCME's expanding consideration of issues relating to classroom assessment, this year's program features several invited sessions related to this important topic. On Saturday, April 14, at 10:35am, *The Past, Present, and Future of Curriculum-Based Measurement* will be discussed, reviewing 30+ years of research in the areas of reading, mathematics, content areas, and writing, and a discussion of future directions and challenges for CBM. On Sunday morning at 10:35am, speakers **Joanna Gorin**, **Margaret Heritage**, and **James Pellegrino** will take on *The Positive Impact of Assessment*, in a conversation about ways that assessment has been a positive impact on teaching and learning as well as ways that it could become a more positive influence in the future. Furthermore, on Monday at 10:35am, we will be joined by **Diego Zapata-Rivera**, **Steven Ferrara**, and **Howard Wainer**, for *We Can Do This: Communicating Information from Educational Assessments*.

Also, as we look forward to the future of assessment, in this year's conference we also take measure of where our field has been. 2018 marks the 25th anniversary of publication of Howard Wainer's *Measurement Problems* article, and we have organized three conference sessions to promote reflection on the past and kick off research for the next 25 years. First, on Saturday, April 14, at 4:05pm we will be joined by several NCME Past Presidents (**Mark Wilson**, **Michael Kolen**, **Suzanne Lane**, and **Laurie Wise**) to ponder where we as a field stand with respect to the original 16 problems, and then on Sunday at 2:45pm and Monday at 12:25pm we will gather again for two forward-looking conversations on unsettled questions and new measurement challenges. The Sunday session panelists include **Li Cai**, **Jimmy de la Torre**, **Chris Han**, **Karen Barton**, and **Alina von Davier**, and Monday will feature **Kathleen Scalise**, **Derek Briggs**, **Andrew Butler**, **Ellen Forte**, and **Sandip Sinharay**. We'd also like to note that **Brian Clauser** has organized sessions to be held on Saturday afternoon focusing on the *History of Measurement from 1950 to the Present*, featuring **Robert Brennan**, **Ronald Hambleton**, **Michael Kane** and **Brent Bridgeman**, and **Michael Bunch**.

For your information, this year's program features 24 workshops, over 50 coordinated sessions, 7 Electronic Board sessions (2 of which are Graduate Student sessions), and over 50 individual paper sessions. Another conference highlight will certainly be the featured session for NCME's Committee on Diversity in Testing, entitled *Insight and Action: Diverse Perspectives on Critical Fairness Issues in Testing*, scheduled for Saturday at 2:15pm. In addition, NCME's 2018 Award Winners will present their award-winning research on Sunday at 4:35pm. If that isn't enough, there's yoga on Saturday morning, receptions nightly, the ever-popular NCME Breakfast and Presidential Address from **NCME President Randy Bennett** on Sunday morning, and the NCME Fitness Run/Walk first thing on Monday.

We would like to take a brief moment to acknowledge the many NCME members who have volunteered over this past year: the vast corps of reviewers who provided us with critically helpful feedback on submitted proposals; our friends and colleagues who were available at all hours to support us with insight, perspectives, and suggestions; and the all of the expert researchers who gracefully accepted our request to serve as discussants for individual paper sessions (they're back this year!). We are exceptionally grateful for your ready willingness to help us throughout this process.

We've been waitin' for you - Enjoy the conference!

April Zenisky and Charlie DePascale
2018 NCME Annual Meeting Co-Chairs

Table of Contents

Here and There and Back Again:
Making Assessment a Stronger Force for
Positive Impact on Teaching and Learning

NCME Officers

President	Randy Bennett *Educational Testing Service*
President Elect	Rebecca Zwick *Educational Testing Service*
Past President	Mark Wilson *UC Berkeley, Berkeley*

NCME Directors

Luz Bay
The College Board

Derek Briggs
University of Colorado

Rose McCallin
CO Dept. of Reg Agencies

Ye Tong
Pearson

Walter Way
The College Board

C Dale Whittington
Shaker Heights (OH) Public Schools

Editors

Journal of Educational Measurement	George Engelhard, *Emory University* Jonathan Templin *University of Kansas*
Educational Measurement Issues and Practice	Dr. Howard Everson *SRI International*
ITEMS Editor	André Rupp *Educational Testing Service*
NCME Book Series Editor	Dr. Brian Clauser *National Board of Medical Examiners*
NCME Newsletter Editor	Megan Welsh *UC - Davis*
NCME Website Editor	Matt Gaertner *WestEd*

2018 Annual Meeting Chairs

Annual Meeting Program Chairs	April Zenisky *University of Massachusetts Amherst*
	Charlie DePascale *National Center for the Improvement of Educational Assessment*
Graduate Student Issues Committee Chair	Masha Bertling *Harvard University*
Training and Development Committee Chair	Amanda Wolkowitz *Alpine Testing Solutions, Inc.*
Fitness Run/Walk Directors	Katherine Furgol Castellano *Educational Testing Service*
	Jill R. van den Heuvel *Alpine Testing Solutions*
	Brain French *Washington State University*

NCME Information Desk

The NCME Information desk is located on the Second Floor Level in the Westin New York at Times Square. Stop by to pick up your badge, as well as your bib number and t-shirt for the fun run and walk. We will also have AERA tote bags and programs this year as well if you registered for both NCME & AERA! It will be open at the following times:

Thursday, April 12 . . . . . 7:30 AM – 4:30 PM
Friday, April 13 . . . . . .7:30 AM - 4:30 PM
Saturday, April 14 . . . . . 8:00 AM – 4:30 PM
Sunday, April 15 . . . . . .10:00 AM – 4:30 PM
Monday, April 16 . . . . . 8:00 AM – 1:00 PM

Proposal Reviewers

Terry Ackerman
Cigdem Alagoz
Alicia Alonzo
Alison Ames
Alvaro Arce
Debbie Bandalos
Bo Bashkov
Kirk Becker
Issac Bejar
Damian Betebenner
Joseph Betts
Laine Bradshaw
Chad Buckendahl
Heather Buzick
Amy Clark
Amanda Clauser
Brian Clauser
Kimberly Colvin
Jenna Copella
Juan D'Brot
Nathan Dadey
Mark Davidson
Jennifer Davis
Laurie Davis
Susan Davis-Becker
Nina Deng
John Donoghue
Jennifer Dunn
Leslie Eastman
Karla Egan
Howard Everson
Tia Fechter
Leah Feuerstahler
Sara Finney
Shameem Gaj
Brian Gong
Chad Gotch
Andrea Gotzman
Sean Gyll
Ron Hambleton
Kristen Huff
Anne Corinne Huggins-Manley
Dan Jurich
Priya Kannan
Lisa Keller
Leslie Keng
Leanne Ketterlin
Min Sung Kim
Seock Ho Kim
Young Kim
G.Gage Kingsbury
Holis Lai
Joni Lakin
Erika Landl
Quinn Lathrop
Sue Lottridge
Xiao Luo
Christine Lyon
Susan Lyons
Ross Markle
Joshua Marland
Krista Mattern
Katie McClarty
James McMillan
Patrick Meyer
Michaelis Michaelides
Rochelle Michel
Dan Mix
Scott Monroe
James Olsen
Tianshu Pan
Thanos Patelis
Marianne Perie
John Poggio
Sonya Powers
Anita Rawls
Mark Raymond
Heather Rickels
Michael Rodriguez
Mary Roduta-Roberts
Yigal Rosen
Louis Roussos
Jonathan Rubright
Andre Rupp
Javarro Russell
Leslie Rutkowski
Christina Schneider
Matthew Schultz
Emily Shaw
Benjamin Shear
Kathleen Sheehan
Mark Shermis
David Shin
Sandip Sinharay
Stephen Sireci
Whitney Smiley
Scott Strickman
Dubravka Svetina
Nathan Thompson
Anna Topczewski
Jill van den Heuvel
Peter van Rijn
Michael Walker
Shudong Wang
Ting Wang
Xi Wang
Xiaolin Wang
Craig Wells
Dale Whittington
Heru Widiatmo
Drew Wiley
Steven Wise
Caroline Wylie
Adam Wyse
Duanli Yan
Hanwook Yoo
Liru Zhang
Ou Zhang
Yue Zhao
Cengiz Zopluoglu

Graduate Student Abstract Reviewers

Christiana Akande
Abeer Alamri
David Alpizar
Ezgi Ayturk
Ella Banda
Yu Bao
Masha Bertling
Tanesia Beverly
Alex Brodersen
Liuhan (Sophie) Cai
Tiago Calico
Luciana Cancado
Delwin Carter
Walter Chason
Michelle Chen
Yi-Chen Chiang
Lilian Chimuma
Dakota Cintron
Lau Clarissa
Veronica Mellado De La Cruz
Rebecca Ellis
Robert Fay
Angela Chinyere
Brittany Flanery
Yanyan Fu
Chenlu Gao
Daria Gerasimova
Stephanie Green
Ying Han
Heather Handy
Ma Hao
Kristen Hochstedt
Liwen Huang
Lim Hwanggyu
Shumin Jing
Tiffany Johnson
Unhee Ju
Stella Kim
Inah Ko
Kevin Krost
Isaac Li
Yujia Li
Ye Lin
Huan Liu
Ren Liu
Ye Ma
Kaiwen Man
James Mason
Stefan Merchant
Ma Mingjia
Zhang Mingqin
Kyle Nickodem
Susan Niessen
Rich Nieto
Mary Norris
Francis O'Donnell
Qianqian Pan
Saemi Park
Justin Paulsen
Duy Pham
Yuxi Qiu
Peter Ramler
Aileen Reid
Jennifer Reimers
Kevin Carl Santos
Deng Sien
Miguel Sorrel
Jordan Sparks
Myrah Stockdale
Meghan Sullivan
Victoria Tanaka
Chen Tian
Lissette Tolentino
Stephanie Underhill
Thao Vo
Jue Wang
Dr. Wu
Qing Xie
Xin Yuan
Matthew Zajic
Jiahui Zhang
Xue Zhang
Xiaying Zheng
Yating Zheng

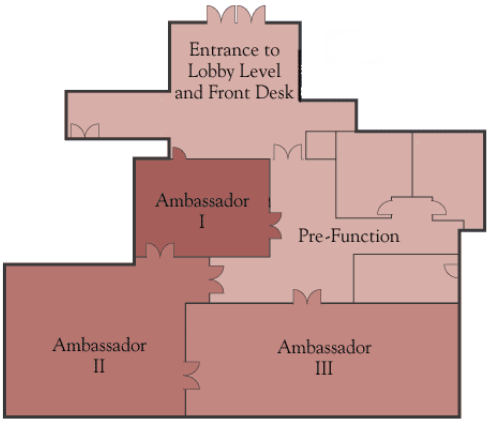
Future Annual Meetings

2019 Annual Meeting
April 4-8
Toronto, Ontario, Canada

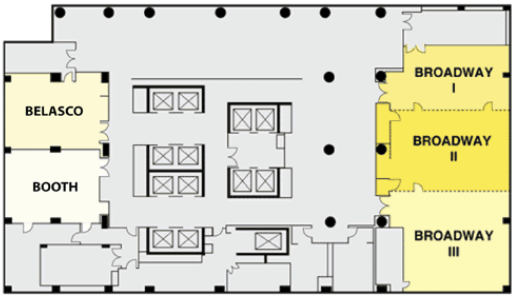
2020 Annual Meeting
April 16-20
San Francisco, CA, USA

Westin at Times Square Floor Plans

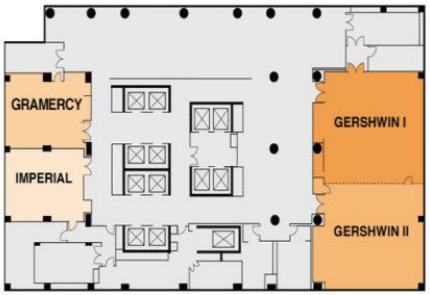
SECOND FLOOR



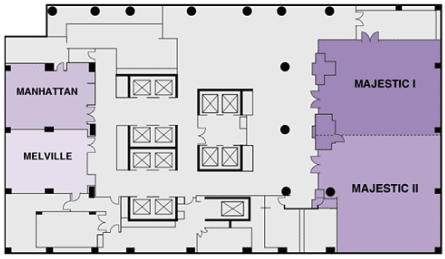
THIRD FLOOR



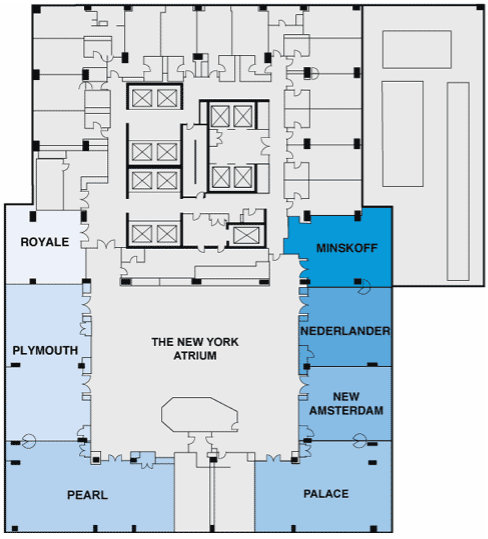
FOURTH FLOOR



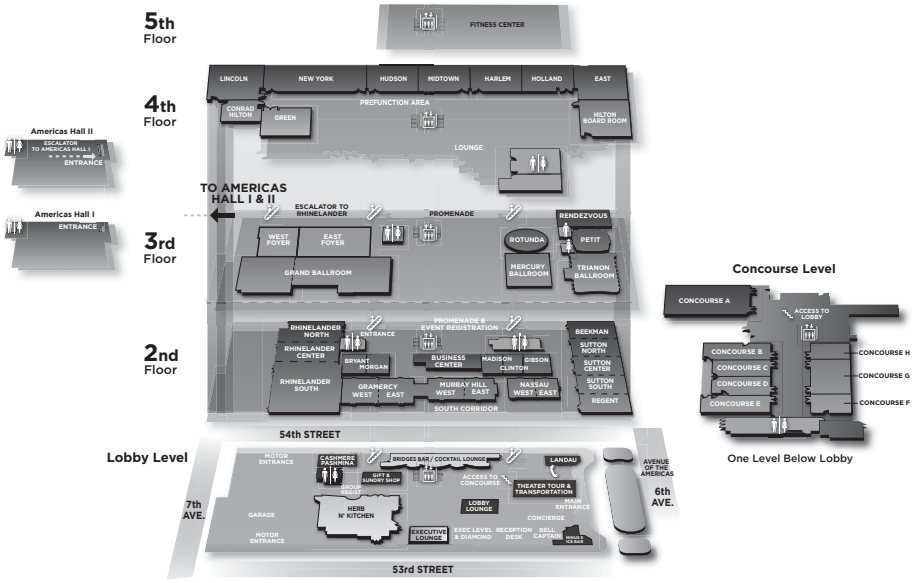
FIFTH FLOOR



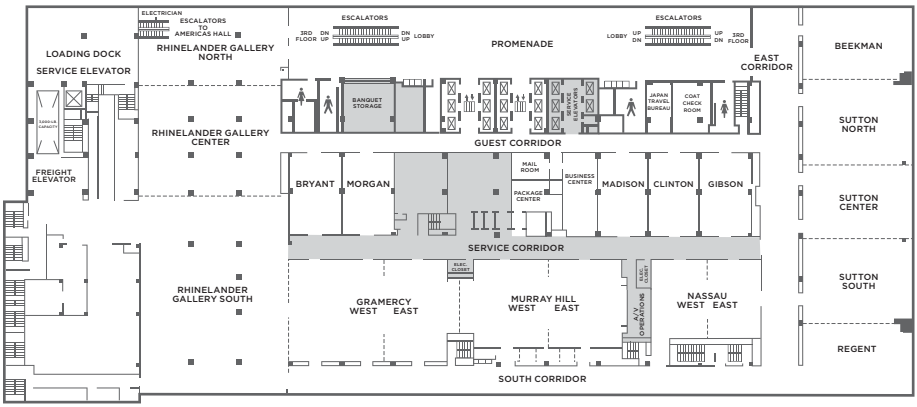
NINTH FLOOR



New York Hilton Midtown Floor Plans



SECOND FLOOR



Pre-Conference Training Sessions

The 2018 Pre-Conference Training Sessions will be held at both the Westin New York at Times Square AND the New York Hilton Midtown on Thursday, April 12. On Friday, April 13, they will just take place at the Westin. All full day sessions will be held from 8:00 AM to 5:00 PM. All half day morning sessions will be held from 8:00 AM to 12:00 PM. All half day sessions will run from 1:00 PM to 5:00 PM.

Onsite registration for the Pre-Conference Training Sessions will be available at the NCME Information Desk at the Westin New York at Times Square for those workshops that still have availability.

Please note that internet connectivity will be available and, where applicable, participants should download the software required prior to the training sessions.

Please ensure to **sign in to all training sessions** you attend, as well as fill out the evaluation after the session. We want to ensure we capture all feedback accordingly so we can provide it to the presenter.

Thursday, April 12, 2018
8:00am-5:00pm, Ambassador III- The Westin, Training Sessions, AA

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

Jimmy de la Torre, The University of Hong Kong
Wenchao Ma, The University of Alabama

This workshop aims to provide participants the necessary practical experience to use cognitive diagnosis models (CDMs) in applied settings. It will also highlight the theoretical underpinnings needed for the proper use of CDMs.

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 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, CDM evaluation at item and test levels, Q-matrix validation, differential item functioning analysis). To ensure 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** and the **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 ability to conduct various CDM analyses using the *GDINA* package. Participants will be requested to bring their laptops for the *GDINA* package hands-on exercises.

Thursday, April 12, 2018
8:00am-5:00pm, Murray Hill East- The Hilton, Training Sessions, BB

Measuring hard-to-measure (noncognitive) skills: Social, emotional, self-management, and beyond

Patrick Kyllonen, Educational Testing Service
Jiyun Zu, Educational Testing Service
Jonas Bertling, Educational Testing Service

This workshop provides training, discussion, and hands-on experience in developing methods for assessing, scoring, and reporting on "hard-to-measure" skills, also known as 21st century, interpersonal and intrapersonal, noncognitive, and social-emotional and self-management skills, for K-12 and higher education. Workshop focuses on (a) reviewing the most important skills based on current research; (b) standard and innovative methods for assessing skills, including self- and others'-ratings, forced-choice (rankings), anchoring vignettes, and situational judgment testing (SJT); (c) cognitive lab approaches for item tryout; (d) classical and item-response theory (IRT) scoring procedures (e.g., 2PL, partial credit, nominal response model); (e) automated test assembly; (f) validation strategies, including rubric and behaviorally anchored rating scale development; (g) reliability from classical test theory, IRT, and generalizability theory; and (h) reporting. Workshop sessions will be organized around item types (e.g., forced-choice, anchoring vignettes). Examples will be drawn from various assessments (e.g., PISA, NAEP, SuccessNavigator, FACETS). There will be hands-on demonstrations using R for scoring anchoring vignettes and SJTs and conducting automated-test-assembly. The workshop is designed for a broad audience of assessment developers and psychometricians, working in applied or research settings. Participants should bring laptops preferably with R and Rstudio installed (but help will be provided if needed).

Thursday, April 12, 2018
8:00am-5:00pm, Clinton Suite- The Hilton, Training Sessions, CC

Techniques and Software for Q-Matrix Estimation and Modeling Learning in Cognitive Diagnosis

Jeffrey Douglas, University of Illinois
Steven Culpepper, University of Illinois
Hua-Hua Chang, University of Illinois
Georgios Fellouris, University of Illinois
Shiyu Wang, University of Georgia
Yinghan Chen, University of Nevada, Reno
Sam Ye, University of Missouri-Kansas City
James Balamuta, University of Illinois
Susu Zhang, University of Illinois

This training session focuses on two areas of cognitive diagnosis that have seen substantial recent developments. The first is estimation of the item-by-attribute matrix, commonly referred to as the Q-matrix. Conditions under which Q can be accurately estimated are considered along with potential implications for test design. Computational techniques will be described and an R software package for fitting the Q-matrix will be discussed along with applications and illustrations. The second part of the session considers how to utilize cognitive diagnosis for modeling learning. A variety of models are introduced combining different measurement models with learning transition models that range from simple first-order hidden Markov models, to models with covariates and continuous latent variables representing learning ability. Computational techniques and methods for assessment of fit will be discussed and illustrated with a dataset concerning an intervention to train spatial reasoning skills. Methods for efficiently and adaptively detecting learning that control for false positives are introduced. Available software for fitting learning models is presented along with applications. The training session is intended for both educational measurement practitioners and researchers. Participants are encouraged to bring a laptop computer with the R software installed.

Thursday, April 12, 2018
8:00am-5:00pm, Murray Hill West- The Hilton, Training Sessions, DD

Using Shiny to create custom psychometric solutions

Joshua Goodman, NCCPA
John Willse, University of North Carolina Greensboro
Reina Chau, NCCPA
Andrew Dallas, NCCPA
Fen Fan, NCCPA

This session explores the role of R and Shiny statistical packages in improving everyday psychometric tasks. Much of the psychometric software programs that are commercially available are offered in a one-size-fits-all format, and thus often lack the flexibility needed within a testing program. Testing organizations with in-house software developers may build custom solutions, but these solutions are often costly and slow to develop. In either scenario, updating a solution to meet the ever-changing needs of a testing program is nontrivial.

Shiny, a freely available R-package implemented within RStudio, is a tool for developing interactive applications which implement customizable solutions to a variety of psychometric activities. Once developed, these apps can be deployed to multiple users either through local installation or an interactive website. Unlike other custom software solutions, Shiny apps are easily developed by psychometric staff with only modest R-programming skills and can be modified quickly and easily as needs evolve. The ability of Shiny to combine rich data collection and display with all of R's analytic power makes it a promising tool for use in a variety of contexts in psychometrics (e.g., item bank analyses, form construction, item analysis, key validation, & standard setting).

Thursday, April 12, 2018
8:00am-12:00pm, Madison- The Hilton, Training Sessions, EE

Computerized Multistage Adaptive Testing: Theory and Applications

Duanli Yan, Educational Testing Service
Alina von Davier, American College Testing
Chris Han, Graduate Management Admission Council

This workshop 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, and its simulations. (Book included)

Thursday, April 12, 2018
8:00am-12:00pm, Ambassador II- The Westin, Training Sessions, FF

Federal Education Policy as a Driver of Assessment Design (1965 to present)

Daniel Lewis, ACT, Inc.
Wesley Bruce, Educational Policy and Assessment Consultant

This training session provides early career measurement professionals (graduate students or DOE staff) with a historical understanding of assessment design and practice in light of federal education accountability policy. The session will provide a historical understanding of current assessment practices resulting from federal education policy to inform next-generation policies and support improved practices.

The presenters trace national educational policy and initiatives, and the associated effects on assessment design and practice, in interactive discussion with participants beginning with President Johnson's original enactment of the Elementary and Secondary Education Act that established accountability requirements that persist today, through each federal administration, to the current Trump administration's policies seeking the decentralization of educational accountability policy.

This session will highlight each administrations' policies that resulted in the push for world-class standards and states' adoption of content standards, the notion of Adequate Yearly Progress in accountability systems, the rise of criterion referenced assessments, testing all students on the same content standards, growth models, wide-spread use of interim and formative assessments, the adoption of common standards (CCSS), adaptive testing, innovative item types, balanced assessment systems, the decentralization of accountability from the federal to state DOEs, and many other topics.

Thursday, April 12, 2018
8:00am-12:00pm, Gibson Suite- The Hilton, Training Sessions, GG

Moving From Paper to Online Assessments: Psychometric, Content, and Classroom Considerations

Susan McNally, Pearson
Ye Tong, Pearson

Online assessments provide a great opportunity for students to interact with content in a more dynamic way. Moving away from paper or to both paper and online assessments requires that we address comparability and consider the best way to assess the content. Creating meaningful online assessments requires clear understanding of the issues faced by psychometricians, content experts, and teachers and students.

In this training session, attendees will learn about the decision-making process for online assessments from both a psychometric and content development point of view. Classroom implications will be discussed based on teaching experience and teacher feedback. Participants will receive instruction around common practices and challenges faced when developing and implementing technology-enhanced items (TEIs). Presenters will demonstrate the most common interaction types to provide attendees with a better understanding. Hands-on experience will be incorporated in the training session around making decisions about assessment formats, item types, and scoring associated with these item types.

Thursday, April 12, 2018
1:00-5:00pm, Gibson Suite- The Hilton, Training Sessions, HH

An Overview of Operational Psychometric Work in Real World

Hyeonjoo Oh, Educational Testing Service
JongPil Kim, ACT
Laura Kramer, The University of Kansas
Jinghua Liu, The Enrollment Management Association
Ye Tong, Pearson

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) discussion session regarding factors that affect operational psychometric activities such as testing mode comparability (paper and pencil test vs. computer based test) and issues with technology enhanced items. 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., ACT, ETS, and Pearson) and University research center developing large-scale assessments will present various topics related to processes in an operational cycle. It is not required to bring a laptop for the session.

Thursday, April 12, 2018
1:00-5:00pm, Madison- The Hilton, Training Sessions, II

Collaborative Solution Design for Educational Measurement Challenges: Not a Spectator Sport

Chad Buckendahl, ACS Ventures, LLC
Ellen Forte, edCount, LLC

Practitioners in educational measurement are frequently confronted with challenges in assessment programs that require solutions that consider psychometric, policy, and business factors. Each of these factors represents a different perspective on what should be prioritized in developing solutions. Although stakeholders for each perspective are often part of the conversation, they may struggle with understanding the alternative viewpoints and how to communicate the important elements of their own. As policies change, programs may be forced to change directions frequently to respond to these competing interests. Developing a better understanding of how these factors interact and where compromises can occur has the potential to lead to better solutions.

This interactive workshop will focus on helping participants develop and practice creative, collaborative problem-solving skills in educational measurement. The structure of the workshop will involve industry leaders iteratively providing information about designing and implementing solutions with this range of factors and then having teams of participants develop proposed solutions based on a common case description. The culmination of the workshop will be having each team present their solution to the panel of industry leaders and other teams as a proposal of their design idea. Panel members will provide reactions to the proposals as feedback.

Thursday, April 12, 2018
1:00-5:00pm, Ambassador II- The Westin, Training Sessions, JJ

Effective Item Writing for Valid Measurement

Anthony Albano, University of Nebraska-Lincoln
Michael Rodriguez, University of Minnesota

In this training session, participants will learn to write and critique high-quality test items by implementing item-writing guidelines and validity frameworks for item development. Educators, researchers, test developers, and other test users are encouraged to participate.

Following the session, participants should be able to: implement empirically-based guidelines in the item writing process; describe procedures for analyzing and validating items; apply item-writing guidelines in the development of their own items; and review items from peers and provide constructive feedback based on adherence to the guidelines. The session will consist of short presentations with small-group and large-group activities. Materials will be contextualized within common testing applications (e.g., classroom assessment, testing with special populations, summative assessment, entrance examination, licensure/certification).

Participants are encouraged to bring a laptop computer, as they will be given access to a web application (http://proola.org) that facilitates collaboration in the item-writing process; those participating in the session in-person and remotely will use the application to create and comment on each other's items online. This practice in item writing will allow participants to demonstrate understanding of what they have learned, and receive feedback on their items from peers and the presenters.

Thursday, April 12, 2018
1:00-5:00pm, Nassau West- The Hilton, Training Sessions, KK

Practical Applications of Vertical Articulation in Standard Setting

Michael Bunch, Measurement Incorporated

This half-day training session, conducted in four modules, takes participants through the key phases of preparing for and conducting vertical articulation of cut scores derived from standard setting for a range of grades. While the presenter will devote most of the first three modules to lecture (each with a Q&A session at the end), the final module will be a mock vertical articulation with the participants playing the role of vertical articulation committee (VAC) members. Participants will receive templates and materials lists for preparing and conducting their own vertical articulations as well as links to additional materials the presenter has uploaded to the NCME ITEMS website.

The primary objective of the session is for participants to acquire the tools and expertise required to plan or conduct vertical articulation, whether they are clients or facilitators. The intended audience will include practitioners, advanced graduate students, and staff of district and state departments of education charged with arranging for or conducting standard setting.

Friday, April 13, 2018
8:00am-5:00pm, Broadway I, Training Sessions, LL

Bayesian Networks in Educational Assessment

Duanli Yan, Educational Testing Service
Diego Zapata, Educational Testing Service
Russell Almond, Florida State University
Roy Levy, Arizona State University

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. (Book is included).

Friday, April 13, 2018
8:00am-5:00pm, Ambassador III, Training Sessions, MM

LNIRT: Joint Modeling of Responses (Accuracy) and Response Times (Speed)

Jean-Paul Fox, University of Twente
Konrad Klotzke, University of Twente

The theoretical foundation of integrating responses and response times in a hierarchical nonlinear and generalized-linear modeling framework is outlined. Next, within an interactive practice session the participants learn how to utilize the free LNIRT R-software to estimate the parameters of interest of the described joint model from a data set which is composed of response times and responses. Attention is paid to the interpretation of item, person and covariance parameters, and specification of explanatory variables for item and person parameters. In a second lecture, tools to evaluate the fit of the joint model will be discussed, including item and person-fit statistics under the joint model. In a practice session, making Bayesian statistical inferences and the validity of inferences made from sequences of Markov Chain Monte Carlo (MCMC) samples and the utility of convergence diagnostics in the given context is discussed. The participants learn how to apply convergence diagnostics to MCMC samples produced by LNIRT using the R-package coda. The training session is aimed at MSc and doctoral students, with a basic knowledge of item response theory and Bayesian statistics, who intend to utilize the LNIRT software to carry out their thesis work or research projects.

Friday, April 13, 2018
8:00am-5:00pm, Broadway II, Training Sessions, NN

Shadow-Test Approach to Adaptive Testing

Bingnan Jiang, ACT, Inc.
Michelle Barrett, ACT, Inc.

The shadow-test approach is not "just another item-selection technique" but an integrated approach to the configuration and management of the entire process of adaptive testing. The purpose of this training session is to (i) introduce the conceptual ideas underlying the approach, (ii) show how it can be used to combine all content, statistical, practical, and logical requirements into a single configuration file, (iii) integrate adaptive calibration of field-test items into operational testing, (iv) use the approach to deliver tests either with a fully adaptive, multistage, linear on-the-fly format or any hybrid version of them, (v) review computational aspects, and (vi) discuss practical implementation issues (dealing with interruptions during testing due to technical glitches, transitioning from fixed-form to adaptive testing, accommodating changes in item pool composition or test specifications, etc.). The session consists of a mixture of lectures, demonstrations, and an opportunity to work with a CAT simulation software program offered to the participants for free. Participants, who are expected to have a medium level of technical knowledge and skills, are encouraged to bring their own laptop computers and item-pool metadata.

Friday, April 13, 2018
8:00am-5:00pm, Broadway III, Training Sessions, OO

Test Equating Methods and Practices

Michael Kolen, The University of Iowa
Robert Brennan, The University of Iowa
Ye Tong, Pearson

The need for equating arises whenever a testing program uses multiple forms of a test that are built to the same specifications. Equating is used to adjust scores on test forms so that scores can be used interchangeably. The goals of the session are for attendees to be able to understand the principles of equating, to conduct equating, and to interpret the results of equating in reasonable ways. The session focuses on conceptual issues. Practical issues are considered.

Friday, April 13, 2018
8:00am-12:00pm, Belasco, Training Sessions, PP

The Stanford Education Data Archive: Using big data to study academic performance

Sean Reardon, Stanford University
Andrew Ho, Harvard University
Benjamin Shear, University of Colorado Boulder
Erin Fahle, Stanford University

The Stanford Education Data Archive (SEDA) is a publicly available dataset based on roughly 300 million standardized test scores from students in U.S. public schools. SEDA now contains average test scores by grade (grades 3-8), year (2009-2015), subject (math and reading/language arts), and subgroup (gender and race/ethnicity) for all school districts in the U.S. Scores from different states, grades, and years are linked to a common national scale, allowing comparisons of student performance over states and time. SEDA was constructed by Sean Reardon (Stanford) and Andrew Ho (Harvard).

This workshop will provide a detailed description SEDA's contents and construction. It will include a description of how test scores are linked to a common scale, the sources and magnitude of uncertainty in the estimates, and appropriate use in descriptive and causal research. Participants will recognize practical applications of test theory, linking, multilevel modeling, and ordinal/nonparametric statistics throughout SEDA's development, in the service of enabling public research.

The workshop will include code, activities, and examples using Stata and R. Participants should bring a laptop with R or Stata, or be prepared to work from raw data using their preferred statistical program.

More information about SEDA is available at http://seda.stanford.edu

Friday, April 13, 2018
8:00am-12:00pm, Majestic I, Training Sessions, QQ

A Visual Introduction to Computerized Adaptive Testing

Yuehmei Chien, Pearson
Ching-Wei Shin, Pearson

The training will provide the essential background information on operational computerized adaptive testing (CAT) with an emphasis on CAT components, CAT simulation, Automated test assembly (ATA), and the multi-stage adaptive testing (MST). Besides the traditional presentation through slides, this training consists of hands-on demonstrations of several CAT key concepts and activities through exercises with visual and interactive tools including a CAT simulator, automated test assembler, MST simulator, and other tools.

Practitioners, researchers, and students are invited to participate. A background in IRT and CAT is recommended but not required. Participants should bring their own laptops, as they will access the tools that were designed to help the participants understand important CAT concepts and tasks and visualize the simulation results. Electronic training materials 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) deeper understanding of the fundamental techniques including simulation, ATA, and MST; 3) an understanding of the costs/benefits/trade-offs of linear vs CAT vs MST test designs; 4) appreciation of the visual techniques used to analyze and present results.

Friday, April 13, 2018
8:00am-12:00pm, Ambassador II, Training Sessions, RR

Applying Test Score Equating Methods Using R

Marie Wiberg, Umeå University
Jorge González, Pontificia Universidad Católica de Chile

The aim of equating is to adjust the test scores on different test forms so that they can be used interchangeably. The goals of the training session are for attendees to be able to understand the principles of equating, to conduct equating, and to interpret the results of equating in reasonable ways. In this training session, different R packages will be used to illustrate how to perform equating when test scores data are collected under different data collection designs. Traditional equating methods, kernel equating methods, IRT equating methods and local equating methods will be illustrated. The main part of the training session is devoted to practical exercises in how to prepare and analyze test score data using different data collection designs and different equating methods. Recent developments in equating are also discussed and examples are provided. Expected audience include researchers, graduate students and practitioners. An introductory statistical background as well as experience in R is recommended but not required.

Friday, April 13, 2018
8:00am-12:00pm, Gershwin II, Training Sessions, SS

Diagnostic Classification Models Part I: Fundamentals

Laine Bradshaw, University of Georgia
Matthew Madison, University of California-Los Angeles

Diagnostic measurement is an emerging field of psychometrics that focuses on providing actionable feedback from multidimensional tests. Part I of this workshop focuses on the foundational theories and psychometric methods of diagnostic classification models (DCMs). More specifically, Part I first provides a conceptual introduction to the models and discusses when the models are well-suited to use. Then Part I introduces the item response functions in a way that builds upon participants' knowledge of general linear models, explains how to interpret both the item parameters and examinee classifications, and illustrates the methods in an empirical setting. Part II of this workshop builds on and extends Part I by introducing more advanced topics of structural models, estimation, and longitudinal versions of the models.

After completing Part I of this workshop, participants will understand the fundamentals of DCMs and be comfortable discussing the functional form and uses of the models. This session is appropriate for graduate students, researchers, and practitioners at the emerging or experienced level. Participants are expected to have only a basic knowledge of statistics and psychometrics to enroll. This session presents both conceptual and technical content. Content will be delivered through lecture and through discussions with the participants and instructors.

Friday, April 13, 2018
1:00-5:00pm, Belasco, Training Sessions, TT

Analyzing NAEP Data Using Plausible Values and Marginal Estimation With AM

Emmanuel Sikali, National Center for Education Statistics
Young Kim, American Institues for Research

Since results from the National Assessment of Education Progress (NAEP) serve as a common metric for all states and select urban districts, many researchers are interested in conducting studies using NAEP data. However, NAEP data pose many challenges for researchers due to its special design features. This class intends to provide analytic strategies and hands-on practices with researchers interested in NAEP data analysis. The class consists of two parts: (1) instructions on the psychometric and sampling designs of NAEP and data analysis strategies required by these design features and (2) the demonstration of NAEP data analysis procedures and hands-on practice. The first part includes the marginal maximum likelihood estimation approach to obtaining scale scores and appropriate variance estimation procedures and the second part includes two approaches to NAEP data analysis, i.e. the plausible values approach and the marginal estimation approach with item response data. The latter is required in analyzing variables not included in the NAEP conditioning models. Demonstrations and hands-on practice will be conducted with a free software program, AM, using a mini-sample public-use NAEP data file released in 2011. Intended participants are researchers, including graduate students, education practitioners, and policy analysts, who are interested in NAEP data analysis.

Friday, April 13, 2018
1:00-5:00pm, Majestic I, Training Sessions, UU

Bayesian Analysis of Response Style IRT Models Using SAS PROC MCMC

Clement Stone, University of Pittsburgh
Brian Leventhal, James Madison University

The training session presents an introduction to Bayesian analysis of IRT models using SAS PROC MCMC with specific applications to response style models. Response style models are designed to account for varying respondent interactions with the response scale of Likert type items in survey research (e.g., tendency to select extreme or midpoint response options). Course content is based on the book, *Bayesian Analysis of IRT Models using SAS PROC MCMC*, and the dissertation research of one presenter. The instructional approach will be one involving lecture and demonstration of the software. Example syntax of models that account for extreme response will be presented and reviewed. Output from PROC MCMC analyzing the models will be discussed. Posterior predictive model checking techniques specific to models accounting for extreme response style will then be explored. The lecture material will be applied in nature and considerable code will be discussed and shared with attendees. The intended audience is graduate students studying measurement as well as researchers interested in the Bayesian paradigm and applications to measurement models. An overall objective of the session is that attendees can extend the examples to their testing applications. Some understanding of SAS programs and SAS procedures is helpful.

Friday, April 13, 2018
1:00-5:00pm, Gershwin II, Training Sessions, VV

Diagnostic Classification Models Part II: Advanced Applications

Matthew Madison, University of California, Los Angeles
Laine Bradshaw, University of Georgia

Diagnostic measurement is an emerging field of psychometrics that focuses on providing actionable feedback from multidimensional tests. Part II of this workshop builds on and extends Part I by providing a more advanced introduction to diagnostic classification models (DCMs). More specifically, Part II focuses on the structural component of DCMs, estimation using R, and recent advancements in longitudinal DCMs.

After completing Part II of this workshop, participants will understand the statistical structure of DCMs, be able to estimate DCMs and interpret software output, and understand how longitudinal DCMs can be applied to assess change in mastery profiles over time.

This session is appropriate for graduate students, researchers, and practitioners at the emerging or experienced level. Participants are expected to have only a basic knowledge of statistics and psychometrics to enroll.

This session presents both conceptual and technical content and also provides hands-on experience for participants to apply what they learn. Material is presented at a technical level when necessary for understanding the models and applying them responsibly. Content will mostly be delivered through lecture, and content will be reinforced using hands-on activities. Instructors will encourage audience participation through questions and allow time for discussions among participants and the instructors.

Friday, April 13, 2018
1:00-5:00pm, Majestic II, Training Sessions, WW

Evidence-Centered Design and Computational Psychometrics Solution for Game/ Simulation-Based Assessments

Jiangang Hao, Educational Testing Service
Robert Mislevy, Educational Testing Service
Alina von Davier, ACTNext
Pak Wong, ACTNext
Kristen DiCerbo, Pearson

Evidence Centered Design (ECD, Mislevy, Steinberg, & Almond, 2003) provides a theoretical framework for designing game/simulation-based assessments. However, in practice, to implement the ECD principles in a particular game or simulation, one must be able to efficiently identify and aggregate the evidence from the complex process data generated as the test taker completes the task. At present, most educational measurement programs do not provide students with rigorous training on how to handle these complex data. In this training session, our goal is to introduce ways to implement ECD in practice, discuss psychometric considerations of modeling process data, and offer hands-on training on how to handle the complex data in terms of data model design, evidence identification and aggregation. We will introduce computational psychometrics (CP; von Davier, 2015; von Davier, Mislevy, & Hao, in progress), which merges data driven approaches with cognitive models to provide a rigorous framework for measuring skills based on process data. The training session is intended for graduate students and educational researchers working on complex items such as games and simulations. Some of the materials used in this workshop are based on an in-progress volume edited by von Davier, Mislevy & Hao (Springer Verlag, expected in 2018).

Friday, April 13, 2018
1:00-5:00pm, Ambassador II, Training Sessions, XX

Landing Your Dream Job for Graduate Students

Deborah J. Harris
Xin Li, ACT, Inc

This training session will address practical topics graduate students in measurement are interested in regarding finding a job and starting a career through four parts. First, what to do now while they are still in school to best prepare for a job, which includes providing suggestions to the questions regarding what types of coursework an employer looks for, how to find a dissertation topic, how to maximize experiences with networking, internships, and volunteering, and what would make a good job talk. Second, how to locate, interview for, and obtain a job, which includes how to find where jobs are, and how to apply for jobs --targeting cover letters, references, and resumes. Third, what to expect in the interview process, which includes job talks, questions to ask, and negotiating an offer. Last, what's next after they have started their first post PhD job, which includes job prospects in the current environment, adjusting to the environment, establishing a career path, balancing work and life, and staying current (i.e., always looking for your next job, either through promotion or transition). The session is interactive, and geared to addressing the participants' questions during the session. Resource materials are provided on all relevant topics.

Saturday, April 14, 2018
8:15-10:15am, Ambassador III, Coordinated Sessions, A1

Are We Entering a New Era for Educational Assessment?

Session Chair: David Conley, University of Oregon
Panelist: Wayne Camara, ACT
Panelist: Margaret Heritage, Independent Consultant
Panelist: Jeff Heyck-Williams, Two Rivers Public Charter School
Panelist: Kristen Huff, Curriculum Associates
Panelist: Scott Marion, Center for Assessment
Panelist: Sloan Presidio, Fairfax County Public Schools

Are we entering a new era for educational assessment? This panel of six leading thinkers, researchers, developers, and practitioners will explore the issue from a variety of perspectives. Some potential questions include: Should the balance tip more toward validity than reliability when judging the quality of an assessment? How can assessment help all students reach high standards? Can non-academic measures help explain variability in student academic performance? What does a multiple-measure system of assessment look like? Can we achieve more coherence and alignment between large-scale assessments and classroom assessment practices? What does a coherent assessment system look like that can support instruction and serve external purposes? Panelists represent a range of perspectives: assessment developer, policy interpreter, researcher, curriculum specialist, system designer, and educational practitioner. The moderator will offer ten principles for a new era system of assessments. Finally, the audience will be invited to respond to questions such as: what are the implications of a new era for educational measurement professionals? What are the new, critical issues to which educational assessment will need to respond? What practices and beliefs are best abandoned? Where do new methods need to be developed? Which might hold the most promise?

Saturday, April 14, 2018
8:15-10:15am, Broadway I, Coordinated Sessions, A2

Advances and Perspectives in Machine Scoring

Session Chair: Mark Shermis, University of Houston-Clear Lake
Session Discussant: Peter Foltz, Pearson

This session examines some of the recent progress in machine scoring relevant to this year's NCME conference theme on teaching and learning. The first paper looks at the application of neural networks to automatically score constructed responses that would not rely so much on the traditional methods of designing and implementing features for automated scoring by hand, but rather using a neural network and an existing corpus of information to develop models for evaluating responses. The second paper studies the application of topic models in formulating possible answer spaces for constructed response items. The third paper proposes a hybrid approach to the scoring of constructed responses. Using this new technique, machine models are modified alongside human scoring until an optimized scoring process has been obtained. The fourth paper also evaluates the impact of creating new models and recommends a set of criteria and rationale for adopting them. The final paper looks at the progress of machine scoring in languages other than English. This involves the scoring of constructed responses in languages ranging from Swedish to Japanese, and the survey suggests better results on packages that are constructed from scratch rather than using extensions of already existing English-language scoring engines.

Deep Learning for Automated Scoring
Aoife Cahill, Educational Testing Service

Topic Model Analysis of Constructed Response Items on a Formative Assessment
Allan Cohen, University of Georgia; Seohyun Kim, University of Georgia; Minho Kwak, University of Georgia; Hye-Jeong Choi, University of Georgia

Continuous Flow: A Hybrid of Human and Automated Scoring
Karen Lochbaum, Pearson

Evaluating Automated Scoring Feature and Modelling Upgrades Relative to Key Criteria
Sue Lottridge, ACT

International Applications of Machine Scoring
Mark Shermis, University of Houston-Clear Lake

Saturday, April 14, 2018
8:15-10:15am, Broadway II, Coordinated Sessions, A3

Understanding, Predicting, and Modifying the Performance of Human Raters

Session Chair: Cathy Wendler, Educational Testing Service
Session Discussant: Walter (Denny) Way, The College Board

While some constructed response (CR) tasks and item types lend themselves to automated scoring, human raters will remain an integral part of CR scoring in the near future. The use of human raters in the scoring process brings with it a number of challenges, since ratings produced by humans can be susceptible to subjectivity and bias. It is crucial to identify and mitigate these challenges and, if needed, modify rater behavior in order to minimize potential rater error, create a cost-effective but accurate scoring system, and ensure the production of valid and reliable scores. The five papers in this session discuss various efforts related to understanding, predicting, and modifying rater behavior along the continuum of CR scoring.

A Conceptual Framework for Examining the Human Essay Rating Process
Edward Wolfe, Educational Testing Service; Diana Ng, Oxford University; Jo-Anne Baird, Oxford University

The Impact of Rater Training on Scoring Accuracy and Efficiency
Ikkyu Choi, Educational Testing Service; Edward Wolfe, Educational Testing Service; Nancy Glazer, Educational Testing Service; Larry Davis, Educational Testing Service; Cathy Wendler, Educational Testing Service

Applying Cognitive Theory to the Human Essay Rating Process
Bridgid Finn, Educational Testing Service; Cathy Wendler, Educational Testing Service

Predictive Rater Models: Implications for Quality Assurance
Isaac Bejar, Educational Testing Service; Chen Li, Educational Testing Service; Daniel McCaffrey, Educational Testing Service

The Impact of Setting Scoring Expectations on Scoring Rates and Accuracy
Cathy Wendler, Educational Testing Service; Nancy Glazer, Educational Testing Service; Brent Bridgeman, Educational Testing Service

Saturday, April 14, 2018
8:15-10:15am, Broadway III, Coordinated Sessions, A4

What Writing Analytics Can Tell Us About Broader Success Outcomes

Session Chair: Jill Burstein, Educational Testing Service
Session Chair: Daniel McCaffrey, Educational Testing Service
Session Discussant: Mya Poe, Northeastern University

Writing is a challenge, especially for at-risk students who may lack the prerequisite writing skills required to persist in U.S. 4-year postsecondary institutions. Educators teaching postsecondary courses that require writing could benefit from a better understanding of writing achievement as a socio-cognitive construct (including writing domain knowledge, general knowledge, and intra- and inter-personal factors). They would also benefit from understanding its role in postsecondary success vis-a-vis retention and college completion. While there is a long tradition of research related to essay writing on standardized tests and college success, typically only the final overall essay score is used for decision-making. This session examines relationships between finer-grained writing analytics and outcomes. We define writing analytics as the study of processes, language use, and genres as they naturally occur in digital educational environments. The five presentations will address these research questions: (1) *How do writing analytics derived from student writing samples relate to measures of broader outcomes*?, and (2) *How might these relationships between writing analytics and broader outcomes inform instruction and assessment to advance student learning*? The presentations will address analytics from a variety of writing genres that reflect on-demand, authentic, and intervention-based writing assignments from 2- and 4-year college students.

Linking Writing Analytics and Broader Cognitive and Intrapersonal Outcomes
Jill Burstein, Educational Testing Service; Daniel McCaffrey, Educational Testing Service; Beata Beigman Klebanov, Educational Testing Service; Guangming Ling, Educational Testing Service

MyReviewers: Understanding How Feedback Supports Writers in Higher Education
Joe Moxley, University of South Florida; Dave Eubanks, University of South Florida

Writing Mentor: Writing Progress Using Self-Regulated Writing Support
Beata Beigman Klebanov, Educational Testing Service; Jill Burstein, Educational Testing Service; Nitin Madnani, Educational Testing Service; Norbert Elliot, University of South Florida

Writing Analytics and Formative Assessment
Charles MacArthur, University of Delaware; Zoi Phillipakos, UNC, Charlotte; Amanda Jennings, University of Delaware

Utility Value and Writing Analytics
Stacy Priniski, University of Wisconsin, Madison; Beata Beigman Klebanov, Educational Testing Service; Jill Burstein, Educational Testing Service; Judith Harackiewicz, University of Wisconsin, Madison; Dustin Thoman, San Diego State University

Saturday, April 14, 2018
8:15-10:15am, Gershwin II, Coordinated Sessions, A5

Developing Simulated Performance Assessments for use in Teacher Licensure

Session Chair: Geoffrey Phelps, Educational Testing Service
Session Discussant: Melissa Margolis, National Board of Medical Examiners

The Educational Testing Service and Teaching Works have collaborated to develop a groundbreaking teacher licensure test assessing three High Leverage Teaching Practices (HLPs) that are essential to safe and effective beginning teaching. Two of the HLPs, Leading a Group Discussion and Eliciting Student Thinking, require the candidate to interact real time in a virtual environment with student avatars. The student avatars are animated by human interaction specialists who can see and hear candidate performances. The four papers in this session present results from a large-scale pilot. Collectively the papers will provide an overview of the assessment and provide examples of task performances, describe training and performance accuracy for interaction specialists, discuss challenges involved in task scoring, and present results from accessibility panels with deaf and blind educators. The session will include opportunities to apply scoring rubrics to actual performances. The discussion will be led by Melissa Margolis from the National Board of Medical Examiners. Dr. Margolis will focus on the challenges of delivering this assessment at scale and its potential for assessing competencies that have not historically been the focus of standardized licensure exams.

Accessibility Challenges, Solutions, and Surprises in Simulated Performance Assessments
Ruth Loew, Educational Testing Service; Eric Hansen, Educational Testing Service; Cara Laitusis, Educational Testing Service

The Standardized Measurement of Beginners' Teaching Competence in a Licensure Exam
Courtney Bell, Educational Testing Service; Francesca Forzani, Teaching Works; Daniel McCaffrey, Educational Testing Service

Training Human Interactors to Deliver Accurate and Standardized of Performances
Sally Gillespie, Educational Testing Service; Morgan Russell, Mursion; Liz Cochran, Educational Testing Service; Carol Forsyth, Educational Testing Service; Christopher Kurzum, Educational Testing Service; Daniel McCaffrey, Educational Testing Service

Scoring the NOTE Pilot: Challenges in Rating Performances of High-Leverage Teaching Practices
Barbara Weren, Educational Testing Service; Nancy Glazer, Educational Testing Service; Alice Sims-Gunzenhauser, Educational Testing Service; Jessica Tierney, Educational Testing Service; Laura Hullinger, Educational Testing Service

Saturday, April 14, 2018
8:15-10:15am, Belasco, Individual Presentations, A6

Exploring Linking Designs

Session Discussant: Michael Walker, Educational Testing Service

A Comparative Study of Three IRT Linking Methods for the Bifactor Model

Kyung Yong Kim, University of North Carolina at Greensboro; Won-Chan Lee, University of Iowa

Under the common-item nonequivalent groups design, the relative performance of the separate, concurrent, and fixed parameter calibration methods for the bifactor model was compared through a simulation study. The three linking methods were evaluated in terms of the recovery of item parameters and accuracy of IRT observed score equating.

Linking Designs within the Context of Continuous Testing

Robert Furter, The American Board of Pediatrics; Saed Qunbar, The University of North Carolina at Greensboro; Andrew Dwyer, The American Board of Pediatrcs

Ongoing, continuous testing poses relatively novel, unique problems for traditional testing practices. This study investigated linking designs within the context of a testing program that includes the same test-takers in consecutive years. Variations of identity, random groups, and anchor test designs were evaluated based on bias, RMSE, and classification accuracy.

Impact of Different Common-Item Designs in Vertical Scaling

John Sessoms, Measured Progress; Louis Roussos, Measured Progress; Wonsuk Kim, Measured Progress

There are different common-item designs for vertical scaling that could affect student ability estimates. We analyzed real vertical data and found meaningful differences in linking constants and student ability estimates across different common-item designs. Test designers likely should carefully consider which common-item design is used.

The Comparison of Several Models for Equating Testlet-Based Test Scores

WenJuan Bu, Beijing Normal University Collaborative Innovation Center of Assessment toward Basic Education Quality; HongBo Wen, Beijing Normal University Collaborative Innovation Center of Assessment toward Basic Education Quality

This study intends to explore the equating accuracy of 2PL model, TRT model and Bi-factor model in Testlet-based tests both in random equivalent groups design and common-item non-equivalent groups design, this study also allows to examine how different scale transformation methods and LID may affect equating using different models.

Evaluation of Sub-score Equating Methods

Rabia Karatoprak Ersen, The University of Iowa; Won-Chan Lee, The University of Iowa

The purpose of the present study is to compare the relative performance of various sub-score equating methods under the common item non-equivalent groups design. Analysis will be conducted through using pseudo-test forms and pseudo groups. The relative performance of the equating methods will be evaluated relative to criterion equating relationships.

Saturday, April 14, 2018
8:15-10:15am, Plymouth, Individual Presentations, A7

Using TIming Data in Innovative Ways

Session Discussant: Pamela Paek, Educational Testing Service

Relationship between Response Time and Characteristics of Items Measuring Teachers' Mathematical Knowledge

Inah Ko, University of Michigan; Patricio Herbst, University of Michigan

This study explores the characteristics of test items using item response time. The finding shows a negative association between response time length and correctness for items requiring memorized knowledge such as definitions, whereas a positive association was found for items requiring both mathematical content knowledge and pedagogical content knowledge.

The Effects of Effort Monitoring with Proctor Notification on Test-Taking Engagement

Steve Wise, NWEA; James Soland, NWEA

When there are no personal consequences associated with performance, test taking engagement represents a serious construct-irrelevant threat to validity. Disengaged test taking often yields scores that are negatively biased and invalid. This study investigates an innovative computer-based test that monitors test-taker effort, and notifies test proctors when students exhibit disengagement.

Exploring Response Time Data to Understand English Learners' Assessment Performance

Mikyung Wolf, Educational Testing Service; Hanwook (Henry) Yoo, Educational Testing Service; Danielle Guzman-Orth, Educational Testing Service

This study examined ELs' response time to better understand their performance on a mathematics assessment across different accommodation conditions (e.g., no accommodation, a glossary version, and a linguistically-modified version). The overall trend by EL membership across the conditions will be presented in relation to the accommodation effects on student performance.

Behavioral indicators of examinee effort on a computer based test

Matthew Grady, American Dental Association; Tina Collier, American Dental Association

This study investigates whether using simple indicators and latent class analysis can provide a valid and reliable method for identifying unmotivated test-takers on a low-stakes assessment. Results suggest that the indicators were successful in identifying one "unmotivated" class and were also more highly correlated with effort than self-reported surveys.

Using the time-signature of items to control testing time in a CAT

Yeow Thum, NWEA; Emily Bo, NWEA

Long testing time subtracts from invaluable instructional time, leads to fatigue and test disengagement, with negative impact on score validity. Tagging items with their time-signatures, defined as predictors of response time, and deploying this information in item-selection procedures help to constraint student testing time in a CAT.

The relationship between effort and accuracy in a computerized large scale assessment

Michalis Michaelides, University of Cyprus; Militsa Ivanova, University of Cyprus; Christiana Nicolaou, Center for Educational Research and Evaluation, Cyprus Ministry of Education and Culture

The study examined the relationship between examinees' test-taking effort and their performance in PISA 2015. The 10% normative threshold method was applied on Science multiple-choice items in the Cyprus sample. Rapid guessers were identified, and their accuracy was lower than accuracy of students engaging in solution-based behavior in most cases.

Saturday, April 14, 2018
8:15-10:15am, Manhattan, Individual Presentations, A8

Evaluating Current and Emerging Psychometric Models and Methods

Session Discussant: Lihua Yao, Defense Personnel Assessment

Routing block requirements for item calibration in multistage testing datasets
Paul Jewsbury, Educational Testing Service; Peter van Rijn, Educational Testing Service Global

In typical multistage tests (MST), later-stage items are only assessed in mutually exclusive subsets of the sample. A sufficient routing block, or set of items taken by the entire sample, was found to be critical in estimating the later-stage item parameters on a common theta metric in pseudo-simulated MST datasets.

Comparison of Bayesian Procedures in Detecting Accuracy-Speed Patterns Indicating Preknowledge
Jin Zhang, ACT Inc.

The effectiveness of two Bayesian procedures is compared in identifying test-takers with item response and response time patterns indicating item preknowledge. The detection rates and the Type I error rates of the procedures are investigated under different conditions where proportions of items and test-takers affected by preknowledge vary.

Subscore Reporting for Double-Coded Items Embedded in Multiple Contexts
Chen Li, University of Maryland, College Park; Hong Jiao, University of Maryland, College Park

This study proposes a two-parameter doubly testlet model with internal restrictions on the item difficulties (MIRID) for subscore reporting. A simulation study is conducted to evaluate the performance of the proposed model in comparison with other models using model parameter recovery accuracy as the criterion.

Comparing Methods of Correcting G-Theory Variance-Component Estimates in the Presence of Nonadditivity
Chih-Kai (Cary) Lin, American Institutes for Research (AIR); Jinming Zhang, University of Illinois at Urbana-Champaign

The current study aims to advance the discussion of nonadditivity in generalizability theory applications and its effects on the estimation precision of variance components. Specifically, the study compared different methods of correcting for bias in variance component estimates in the presence of nonadditivity.

The Dilemma of Interaction in a One-Facet Generalizability Model
Jinming Zhang, University of Illinois at Urbana-Champaign; Chih-Kai (Cary) Lin, American Institutes for Research

Interaction in a one-facet model is considered a part of compound residual in generalizability theory. This study shows, however, that if interaction exists in data, the interaction cannot be treated as a part of residual, or generalizability models are no longer appropriate for the data. A nonadditive solution is proposed.

Using Concerto for Experimental Research on CAT: Lessons Learned
Yi Zheng, Arizona State University

This presentation will share the lessons learned from configuring and using Concerto, an open-source platform for developing and delivering computerized adaptive tests (CAT). Aspects to be discussed include installing and configuring Concerto, developing CAT using the platform, and running a pilot experimental study using the developed CATs.

Saturday, April 14, 2018
8:15-10:15am, Ambassador II, Individual Presentations, A9

Setting Performance Standards: New Contexts and Approaches

Session Discussant: Susan Davis-Becker, ACS Ventures

Predicting College Readiness in STEM: Establishing STEM Readiness Benchmarks

Heather Rickels, University of Iowa, Iowa Testing Programs; Catherine Welch, University of Iowa, Iowa Testing Programs; Stephen Dunbar, University of Iowa, Iowa Testing Programs

This study established STEM Readiness Benchmarks on a state achievement test for typical first-year STEM courses. These benchmarks were compared to general college readiness benchmarks. Results indicate that STEM coursework is more demanding and students need to be better prepared academically if planning to pursue a STEM degree.

Setting New Standard for Tests in Transition

Ourania Rotou, Educational Testing Service; Han Por, Educational Testing Service

An alternative approach to standard setting procedures is used to set new standards for tests in transition. Score correspondence is evaluated for 12 combinations of interest. The correlation between the new and unchanged parts of the revised test is an important factor in the quality of score correspondence.

Divide and Conquer: An Angoff Modification

Jerome Clauser, American Board of Internal Medicine; Kelly Foelber, American Board of Internal Medicine

This study examines how a modification to the Angoff method affects the validity of passing scores. Experimental data were used to examine the impact of having content experts sort items by difficulty prior to rating. The results show that the internal consistency improved for all six exams in the study.

Interval Validation Method: An Investigation of Large Pool Sizes and Content Balance

William Insko, Houghton Mifflin Harcourt

The Interval Validation Method for setting achievement level standards is specifically designed for assessments with large item pools. The present study uses simulation techniques to validate the use of the method with very large pools while assuring content balance. Several recommendations for controlling content coverage for Exemplar items are discussed.

An Experimental Evaluation of Structured Feedback in Angoff Standard Setting

Janet Mee, National Board of Medical Examiners; Brian Clauser, National Board of Medical Examiners; Melissa Margolis, PhD; Peter Baldwin, National Board of Medical Examiners; Marcia Winward, National Board of Medical Examiners

Research suggests that content experts have difficulty making the estimates required in Angoff standard setting. This research experimentally evaluated a training procedure designed to improve judge accuracy. One group viewed performance data; the other group received no feedback. The results do not suggest that the feedback improves subsequent accuracy.

Achievement Level Description Validation: Starting from Scratch with Student Performance Data

Pamela Kaliski, College Board; Lei Wan, College Board; Rosemary Reshetar, College Board; Leigh Abts, University of Maryland; Jennifer Kouo, Towson University

SMEs who participate in standard settings often participate in ALD validation by evaluating alignment between empirical data and standard setting ALDs. In this study, SMEs would not be standard setting participants, and they will write ALDs from scratch. This rigorous approach could offer rich feedback to inform potential revisions.

Saturday, April 14,2018
10:35 – 12:05, Ambassador 3, Invited Session, B1

The Past, Present, and Future of Curriculum-Based Measurement

Session Chair: Kristen McMaster, University of Minnesota
Session Discussion: Kristen McMaster

A session providing an overview and history of Curriculum-Based Measurement (CBM), a review of the 30+ years of research in the areas of reading, mathematics, content areas, and writing, and a discussion of future directions and challenges for CBM.

CBM was developed by Stan Deno and his colleagues at the University of Minnesota in the 1980s. As described by Deno in an article on the development of CBM: *Curriculum-based Measurement (CBM) is an approach to measuring the academic growth of individual students. The essential purpose of CBM is to aide teachers in evaluating the effectiveness of the instruction they are providing to individual students. Early research focused on testing the utility and effectiveness of CBM for increasing the achievement of special education for students with learning disabilities. Extensions of CBM research now address a broad range of educational issues in both special and general education with different populations and in new curriculum domains.*

A persisting alternative: Overview and history of CBM
Kristen McMaster, University of Minnesota

Extending CBM for Written Expression to Emerging Writers
Kristen Ritchey, University of Delaware; David Coker, University of Delaware

Science is Golden: Using CBM in Content Areas
John Hosp, University of Massachusetts Amherst

Pre-service and novice educators' CBM graph interpretation skills and potential implications for training
Dana Wagner, Minnesota State University Mankato

Saturday, April 14, 2018
10:35am-12:05pm, Broadway I, Coordinated Sessions, B2

Challenges and Opportunities on International Higher Education Admission Practices

Session Chair: Maria Elena Oliveri, Educational Testing Service
Session Discussant: Cathy Wendler, Educational Testing Service

The coordinated session will discuss key challenges and current and future-looking opportunities related to broadening postsecondary admissions criteria to augment the predictive validity of test score use, inform the assessment of alternative outcomes including later job performance, and increase access to diverse populations (e.g., as defined by race/ethnicity, gender, or socioeconomic status) to support fairness and advance opportunity to learn. The presentations will be from international scholars (Cyprus, Israel, U.S., Holland, and Sweden). From the national perspective of each participant, the presenters will describe lessons learned in admissions decision-making practices to shed light on considerations related to diversity in measuring achievement using various (a) background factors in score use and interpretation, (b) criteria--such as previous classwork, grades, test scores, personal experience, and recommendation letters, and (c) perspectives in decision-making processes. The session will provide an overview of international postsecondary institutional practices used to inform admissions decisions. The goal is to inform current and innovative approaches to improve the relevance, utility, and consequences of higher education assessments for diverse stakeholders including policymakers, assessment developers, and psychometricians who embrace common goals of promoting a fair and equitable admissions process.

Distributed and Local Assessment Paradigms: Can They Co-Exist in Symbiotic Ways?
Robert Mislevy, Educational Testing Service; Maria Elena Oliveri, Educational Testing Service; Norbert Elliot, USF

Using Intrapersonal and Interpersonal Skills Assessments in Admission Procedures: An International Perspective
Rob Meijer, University of Groningen, The Netherlands; Susan Niessen, University of Groningen, The Netherlands

Group Differences in Selection to Swedish Postsecondary: Implications for Fairness and Equity
Christina Wikstrom, Umea University; Magnus Wikstrom, Umea University

Changes and Revisions in Higher Education Admissions in Israel
Avi Allalouf, National InstituteTesting & Evaluation; Naomi Gafni, National Institute Testing & Evaluation

Perceived Fairness and Equity in Developing and Using Admissions Assessments in Cyprus
Elena Papanastasiou, University of Nicosia; Michalis Michaelides, University of Cyprus

Saturday, April 14, 2018
10:35am-12:05pm, Broadway II, Coordinated Sessions, B3

Validity Considerations for New Data in Performance Learning and Assessment

Session Chair: Saskia Wools, Cito
Session Discussant: James Pellegrino, University of Illionois

This symposium will discuss the validity implications of new innovations in assessment – including computer-based, dynamic multimodal forms of assessment, makerspace, games and simulations. These innovations significantly expand the information base that can be used to make judgements about assessment performance, learning and validity. In this symposium, we will present and discuss the state-of-the-art, and how new sources of 'process data' and 'para-data' such as key strokes and clicks, and micro-analytic studies of interaction are being integrated into learning and assessment. These new developments challenge us to re-think what counts as data and how these new sources of information are integrated into a validity frameworks. We will ask: 'What opportunities and threats do these innovations in data create for validity theory and practice?'

The presenters and discussant in this symposium include leading researchers in performance assessments and validity studies. In the symposium we aim to stimulate, inform and challenge the audience to consider the radical implications of new technologies and new data practices for assessment design, and how we think of validity. The papers will locate these examples theoretically with discussions of the state of the art drawn from recent publications in learning analytics, assessment research and validity theory.

Assessment and Validity In-Vivo
Bryan Maddox, University of East Anglia; Bruno Zumbo, University of British Columbia

Computational Psychometrics: Conceptual Model for the Validity of Learning and Assessment Systems
Alina von Davier, ACT Next

Bridging Gaps between Multimodal Data and Ecologically Valid Assessment of Complex Skills
Saad Khan, Educational Testing Service

Performance learning and assessment within an argument-based approach
Saskia Wools, Cito

Saturday, April 14, 2018
10:35am-12:05pm, Broadway III, Coordinated Sessions, B4

Experimental Design within a Survey Assessment: Learning from NAEP Digital Transition

Session Chair: Yue Jia, Educational Testing Service
Session Discussant: Peggy Carr, NCES
Session Discussant: Matthew Johnson, Department of Human Development

NAEP, the "Nation's report Card", is principally concerned with the ability to track performance of U.S. populations across time. NAEP's transition from paper-based to digital-based administration, conducted between 2015 and 2017, is particularly complicated in terms of design principles (i.e., how items are presented, what kinds of evidence can be collected) and assessment methodologies (i.e., how bridge studies are designed and analysis methodologies applied in a group score environment). NAEP is uniquely able to engage in careful study of differences across modes, yielding rich insights into possible differentiating factors that might otherwise remain uncovered, and developing deep understanding on how test scores across different delivery modes is to be interpreted nationally, and across urban districts and states.

The collection of papers provide insight into what we have learned from the NAEP mode transition studies from a technical and methodological perspective, along with empirical results. As many assessments are transitioning to digital delivery, score comparability are of relevant for practitioners, policy makers, and researchers alike. With suitable design in place, we demonstrate that the knowledge developed in NAEP give other assessments, particularly the K-12 programs, first-hand knowledge as they consider their own digital transition.

Digital Transition of Group Score Survey Assessments: Design Rules and Lessons Learned
Yue Jia, Educational Testing Service; Andreas Oranje, Educational Testing Service

Experimental Design within Survey – Sample Design for the Transitional NAEP Assessments
Tien-Huan Lin, WESTAT; Natalia Weil, WESTAT; Keith Rust, WESTAT

Evaluation and Modelling of Impact of Paper-to-digital Transadaptation on a Measurement Instrument
Paul Jewsbury, Educational Testing Service; Nuo Xi, Educational Testing Service; Meng Wu, Educational Testing Service

Are Group Scores Comparable? –Subgroup Results from 2017 NAEP DBA Transition
Nuo Xi, Educational Testing Service; Paul Jewsbury, Educational Testing Service; Meng Wu, Educational Testing Service

Saturday, April 14, 2018
10:35am-12:05pm, Gershwin II, Coordinated Sessions, B5

Digitally Simulated Science Laboratory Assessments: Differential Approaches for Analyzing Log File Data

Session Chair: Man-Wai Chu, University of Calgary
Session Discussant: André Rupp, Educational Testing Services

Digitally simulated laboratory assessments (DSLA) are becoming popular supplements to traditional classroom performance measures because of their potential to non-intrusively collect student process data and enhance students' socio-emotional experiences during learning (Chu, 2017). Log files, which reflect students' actions during DSLAs, provide evidence of both content knowledge and process skills. However, there are challenges associated with the analyses of these log files. The papers presented during this symposium demonstrate different analyses and strategies for overcoming challenges that could be used to extract evidence of students' scientific understanding, as well as to explore students' socio-emotional experiences while completing DSLAs. The results of these studies indicate that analysing log files using a simple rubric marking system, adaptive subspace self-organizing map, and long short term memory cluster approach show promise in providing evidence to support claims of student scientific understanding that would not otherwise be easily collected using traditional laboratory assessments. Furthermore, the final paper builds on the results from these analyses by incorporating socio-emotional variables into the interpretation of students' cognitive outcomes. Collectively, results from the four studies provide directions for future analytical research of log files so that substantive claims of student performance and processing skills may be made using DSLAs.

The use of Digitally Simulated Laboratories as Educational Assessment Tools
Man-Wai Chu, University of Calgary; Jacqueline Leighton, Unviersity of Alberta; Qi Guo, University of Alberta; Ying Cui, University of Alberta

Logdata Feature Extraction with Adaptive-Subspace Self-Organizing Map: A Neutral Network Approach
Ying Cui, Unviersity of Alberta; Qi Guo, University of Alberta; Jacqueline Leighton, University of Alberta; Man-Wai Chu, University of Calgary

Use Bayesian Networks to Analyze Logfile data and Compare with TRESim Results
Qi Guo, Unviersity of Alberta; Ying Cui, University of Alberta; Man-Wai Chu, University of Calgary; Jacqueline Leighton, University of Alberta

Adding Value to Diagnostic Test-Based Inferences: The Case for Socio-Emotional Inputs
Jacqueline Leighton, University of Alberta; Man-Wai Chu, University of Calgary; Ying Cui, University of Alberta; Qi Guo, University of Alberta

Saturday, April 14, 2018
10:35am - 12:05pm, Belasco, Individual Presentations, B6

Estimating Parameters in an Adaptive Context

Session Discussant: Daniel Bolt, University of Wisconsin

The Impact of Item Parameter Drift in Computer Adaptive Testing

Sien Deng, University of Wisconsin-Madison; Meichu Fan, ACT, Inc.; Qing Yi, ACT, Inc.

This study investigates the effect of item parameter drift (IPD) in computer adaptive testing (CAT) for two item selection methods, especially aims to evaluate and compare the performance of these two item selection methods on measurement precision, classification accuracy, and test efficiency under varying IPD conditions.

Comparison of Calibration and Drift Detection Methods under Multistage Testing

Liuhan Cai, University of Nebraska-Lincoln; Louis Roussos, Measured Progress; Xi Wang, Measured Progress

As a testing program transitions from a linear design to a pre-equated multistage test (MST), it is important to evaluate item parameter drift from the pre-equated model. This study investigates the performance of several item calibration and drift detection methods implemented for this purpose.

Using Design Information in Item Parameter Estimation with Multistage Testing

Ru Lu, Educational Testing Service; Helena Jia, Educational Testing service; Meng Wu, Educational Testing Service

The objective of this study is to investigate whether using the multistage testing (MST) design information can improve item parameter estimation based on MST data. We choose a two-stage MST design to illustrate the estimation procedure and compare the results with the practice of without using design information.

Pedagogical Applications for Estimation and Scoring in Item Response Theory

Metin Bulus, University of Missouri; Wes Bonifay, University of Missouri

Research in teaching advanced measurement topics has been a scarce. Furthermore, there are limited applications to assist researcher to teach hard-to-grasp concepts such as estimation. We provide interactive tools for students, instructors or researcher to explore or to investigate fundamental concepts in item response theory.

Evaluating Item Parameter Drift for Vertical Linked Large-Scale Computer Adaptive Tests

Tao Jiang, American Institutes for Research; Xiaodong Hou, American Institutes for Research; Stephan Ahadi, American Institutes for Research

The paper describes a fit statistic to implement in large-scale adaptive assessments, making it possible to monitor item performance throughout assessment windows. We compare its performance with a recalibration approach, examine item attributes being more susceptible to drift, and identify steps to be taken for items with evidence of drift.

Saturday, April 14, 2018
10:35am - 12:05pm, Plymouth, Individual Presentations, B7

Diagnostic Classification Models: Challenges and Opportunities

Session Discussant: Mark Hansen, UCLA

Incorporating Item Features into Diagnostic Classification Models
Manqian Liao, University of Maryland College Park; Hong Jiao, University of Maryland College Park

This study proposes a diagnostic classification model(DCM) with item features incorporated. Specifically, the item features are linked to the guessing/slipping probabilities in the deterministic-input-noisy-and-gate(DINA) model for explanatory purposes. It potentially provides guidance on item writing for cognitive diagnosis. Additionally, the item parameter estimates are more precise with additional information incorporated.

Assessing Academic Growth in the Diagnostic Classification Model Framework
Qianqian Pan, Achievement and Assessment Institute, University of Kansas; Neli Kingston, University of Kansas

States require abstract Assessing academic growth. A simulation study is conducted to investigate how the number of common items per attribute, the number of total items per attribute, and types of Diagnostic Classification Models (DCMs) could affect measures of academic growth in a variety of conditions using DCMs.

Item Influence Measures for Diagnostic Classification Models
Matthew Madison, University of California, Los Angeles

Because diagnostic classification models coarsely classify examinees into groups, they can achieve high reliability with fewer items (e.g., <10). One unintended consequence of constructing tests with fewer items is that an item may have disproportionate influence on examinee classifications. This study introduces measures of item influence and examines their performance.

A Diagnostic Classification Model for Polytomous Attributes
Yu Bao, University of Georgia; Laine Bradshaw, University of Georgia

Most diagnostic classification models (DCMs) provide dichotomous feedback about students' mastery and non-mastery levels. In some educational scenarios, further delineating mastery categories into 3, 4 or 5 levels may be useful for tailoring instruction. We propose a polytomous DCM (PDCM) that classifies students into more than two mastery levels.

Designing Field Tests for Multidimensional Classification Models
Selay Zor, University of Georgia; Laine Bradshaw, University of Georgia

Field testing is an essential step in creating assessments. Multidimensional assessments introduce new challenges for field testing. We investigate the impact of planned missing data, due to different field test designs, on the estimation accuracy for diagnostic classification models (DCMs). Results provide needed guidelines for designing DCM-based field tests.

Saturday, April 14, 2018
10:35am - 12:05pm, Manhattan, Individual Presentations, B8

Validating Assessments for Particular Uses

Session Discussant: Drew Wiley, ACS Ventures

The GRE as a Predictor of Law Student Success: A Generalizability Study

David Klieger, Educational Testing Service; Brent Bridgeman, Educational Testing Service; Richard Tannenbaum, Educational Testing Service; Frederick Cline, Educational Testing Service

After law schools identified several non-psychometric benefits to using the GRE test in admissions, they asked whether using GRE scores also would be psychometrically valid generally. We therefore undertook a generalizability criterion-related validity study with 23 public and private U.S. law schools of varying selectivity levels, geographic locations, and sizes.

Test Fairness and Validity in Measuring Domain-specific Knowledge and Understanding in Economics

Jasmin Schlax, Johannes Gutenberg-University Mainz; Olga Zlatkin-Troitschanskaia, Johannes Gutenberg-University Mainz; Carla Kühling-Thees, Johannes Gutenberg-University Mainz; Judith Jitomirski, Humboldt-University; Roland Happ, Johannes Gutenberg-University Mainz; Sebastian Brückner, Johannes Gutenberg-University Mainz; Manuel Förster, Johannes Gutenberg-University Mainz; Hans Pant, Humboldt-University Berlin

Deficits in test fairness impair valid score interpretations of assessed student learning outcomes. When validating an economic knowledge test, in a sample of 9,055 first-year students from 54 German universities, we found non-native speakers showed significantly more non-responses than native speakers. Imputation procedures meant to minimize bias will be discussed.

Rater Fatigue in a Medical Certification Performance Exam

Tara McNaughton, Measurement Incorporated; Carol Myford, University of Illinois at Chicago

We identified differential accuracy, consistency, and central tendency across days and sessions (AM/PM) as potential signs of rater fatigue in a certification exam. Accuracy revealed significant differences across days and sessions while central tendency differences were only discernible across sessions. There was no evidence of increased inconsistency.

A Comparison of Subject Matter Experts' Perceptions and Job Analysis Surveys

Adam Wyse, The American Registry of Radiologic Technologists; Ben Babcock, The American Registry of Radiologic Technologists

Two common approaches for performing job analysis in credentialing programs are committee based methods that rely solely on subject matter experts' judgments and task inventory surveys. This study evaluates how well subject matter experts' perceptions coincide with task inventory survey results for three credentialing programs.

Differential Prediction by Gender in Performance-sampling Tests for College Admissions

Susan Niessen, University of Groningen; Rob Meijer, University of Groningen; Jorge Tendeiro, University of Groningen

This study investigated the potential of performance-sampling admission tests to reduce differential prediction by gender, using a Bayesian approach. Performance sampling did not completely eliminate differential prediction, but when it was detected the effect sizes were small. In addition, differential prediction effect sizes were smaller for more comprehensive curriculum samples.

Saturday, April 14, 2018
10:35am - 12:05pm, Ambassador II, Individual Presentations, B9

Perspectives on Response Modeling

Session Discussant: Edward Wolfe, Educational Testing Service

A gibbs sampler for higher-order item response model

Zhihui Fu, Northeast Normal University , Shenyang Normal University; Jian Tao, Northeast Normal University; Xue Zhang, Northeast Normal University

Based on an efficient data augmentation scheme, using the Gibbs sampler, we propose a Bayesian procedure to estimate higher-order 3PLM. With the introduction of two latent variables, the full conditional distributions are tractable, and the Gibbs sampling is consequently easy to implement.

Impact of model misspecification on a nonlinear reliability for categorical responses

Hye-Jeong Choi, University of Georgia

The purpose of this study is to investigate the impacts of model misspecification on a nonlinear reliability coefficient for categorical responses. A brief description of the nonlinear reliability coefficient will be presented followed by a simulation study and empirical data analyses with three data sets.

A Bifactor Approach to Modeling Dependencies between Response Time and Accuracy

Haiqin Chen, American Dental Association; Paul De Boeck, The Ohio State University; Matthew Grady, American Dental Association; Chien-Lin Yang, American Dental Association; David Waldschmidt, American Dental Association

A bifactor model is proposed to capture dependencies between response time and accuracy across test content domains. These dependencies were found to be positively correlated with item difficulties for all domains considered. As items became easier, the local dependence was increasingly negative.

Measurement of digital competence: how to model and report

Perman Gochyyev, University of California, Berkeley; Fazilat Siddiq, Nordic Institute for Studies in Innovation, Research and Education

Multidimensional item response models are used for analyzing multicomponent data. There are closely related models that deal with data from tests measuring multiple dimensions, such as bifactor and second-order models. We consider all three of these models in the context of measurement of the digital competence.

Two-Level Alternating Direction Model for Polytomous Items Violating Local Independence

Igor Himelfarb, The National Board of Chiropractic Examiners; Bruce Shotts, The National Board of Chiropractic Examiners; Guoliang Fang, Lincoln University

In this paper, we propose a two-level unidimensional TRT estimation algorithm which inherits the advantages of approaches at different levels. The model accounts for local independence in polytomous testlet items. The algorithm is developed and tested with simulations and real test data.

Saturday, April 14, 2018
10:35am - 12:05pm, Gershwin I, Individual Presentations, B10

Electronic Board Session 1

Electronic Board #1

Evaluating mode comparability in early elementary grades

Ye Lin, University of Iowa; Catherine Welch, University of Iowa; Stephen Dunbar, University of Iowa

Mode comparability in this study includes a consideration of test-level invariance in the MANOVA framework, as well as the item-level and construct-level invariance as conceived with DIF and CFA framework. Multiple methodologies help to develop a complete picture of comparability between modes in early elementary grades.

Electronic Board #2

Can Student Attitude Toward STEM-PBL Predict Their Academic Achievement?

Mahnaz Shojaee, University of Alberta, Educational Psychology, "Measurement, Evaluation & Cognition"; Ying Cui, University of Alberta, Educational Psychology, Measurement, Evaluation & Cognition; Mehrdad Shahidi, Mount Saint Vincent University

Focusing on the degree to which the students' attitude toward science, technology, engineering, math and project-based learning (STEM-PBL) may predict academic achievement, this study also investigates the relationship between student attitude toward STEM-PBL and their demographic variables including GPA, gender, parents' education and the time of usage of technology.

Electronic Board #3

Improving College and Career Readiness: The SAT Suite of Assessments

Daria Gerasimova, George Mason University; Scott Marion, The National Center for the Improvement of Educational Assessment

Test developers and others are creating tools to help high schools improve their students' levels of college and career readiness. This study explores how districts use one set of tools, College Board's SAT Suite, and how it influences the ways in which they support the readiness levels of their students.

Electronic Board #4

A Mixture IRT Model of a Statewide Kindergarten Entry Formative Assessment

Do-Hong Kim, Augusta University; Richard Lambert, University of North Carolina at Charlotte

This study used the mixture IRT model to determine whether there exist two or more distinct sub-populations with respect to the measurement of school readiness. The data for this study come from a Kindergarten Electronic Portfolio Assessment in the Statewide Kindergarten Entrance Assessment in 2016. Three distinct latent classes emerged.

Electronic Board #5

Developing Scoring Guidelines for Performance Assessments Using Expert Judgments: An Exploratory Study

Ulana Luciw-Dubas, National Board of Medical Examiners; Polina Harik, National Board of Medical Examiners; Monica Cuddy, National Board of Medical Examiners; Constance Murray, National Board of Medical Examiners; Cara Artman, National Board of Medical Examiners

Performance assessments are effective for evaluating competency in complex constructs, such as patient management, professionalism and communication. However, developing scoring guidelines for performance assessments remains challenging. We discuss one approach to developing scoring guidelines for a high-stakes performance assessment of physicians' patient management skills and compare it with other methods.

Electronic Board #6

Examining Effectiveness of Double Linking in Population Invariance

Yan Huo, Educational Testing Service

The purpose of this study is to investigate the impact of a test form linking design (single vs. double) on the assumption of population invariance. To this aim, four factors are considered in a simulation study: single/double linking, group heterogeneity, anchor set length, and form difficulty difference.

Electronic Board #7

Using Cross-Classified IRT Models to Investigate Test Performance

Youhua Wei Wei, Educational Testing Service

It is common practice in IRT to consider item effect as fixed and person effect as random. This study will use cross-classified random effects two-parameter IRT models with person and item covariates to investigate their impact on the performance of a test (e.g. item difficulty, discrimination, and person ability).

Electronic Board #8

Exploring Effectiveness of Sequential Monitoring Procedures in Detecting Suspicious Examinees in CAT

Nooree Huh, ACT, Inc.; Qing Xie, The University of Iowa/ ACT, Inc.; Chunyan Liu, ACT, Inc.; Chi-Yu Huang, ACT, Inc.

This study examines the effectiveness of sequential item monitoring procedures that are used to identify breached items in detecting possible cheaters. Different factors are considered including the characteristic of cheater occurrences, the percentage of cheaters, the difficulties of compromised items, the percentage of compromised items, and moving sample size.

Electronic Board #9

A comparison among four item-level fits for the DINA model

Xiaojian Sun, Beijing Normal University; Tao Xin, Beijing Normal University

Results of simulation study show that: (1) PD^* and Q_1^* lead to better Type-I error rate than RMSEA and MAD in most conditions regardless of type of misspecification; (2) with sample size and proportion of misfit item increased, Type-I error rate for *Q*-matrix and model misspecification increased.

Electronic Board #11

Measuring Food Security Using a Bifactor Model in Households with Children

Victoria Tanaka, The University of Georgia; George Engelhard, Jr., The University of Georgia; Matthew Rabbitt, The United States Department of Agriculture Economic Research Service; J. Jennings, The National Board of Examiners in Optometry

Research suggests the U.S. Household Food Security Scale exhibits multidimensionality when children are present in the household. We explore multidimensionality using bifactor models and data for low-income households with children. The data suggest a general dimension captures overall household food security, and specific dimensions capture adult and child food security.

Electronic Board #12

A Cognitive Diagnosis Model for Ordinal Data

Charles Iaconangelo, Pharmerit International; Daniel Serrano, Pharmerit International

To address the lack of cognitive diagnosis models (CDMs) for ordinal data, a new CDM is proposed that assigns examines to latent classes via multinomial item responses. An EM algorithm for item parameter estimation is developed; a simulation study illustrates the viability of the model.

Electronic Board #13

CutScore: A Shiny App for the Cut-Score Operating Function

Christopher Runyon, The University of Texas at Austin; Irina Grabovsky, National Board of Medical Examiners

We introduce "CutScore," an app made in the R Shiny environment that implements the Cut-Score Operating Function method for standard setting. The application allows one to estimate the size of classification errors for various possible cut-scores and to calculate the optimal value of the cut-score which minimizes the classification error.

Electronic Board #14

Stratified Diagnostic Methods for Rasch Based Computerized Adaptive Test

Qinjun Wang, University of Minnesota - Twin Cities; Michael Rodriguez, University of Minnesota - Twin Cities; Alisha Hollman, University of Minnesota - Twin Cities; Scott McConnell, University of Minnesota - Twin Cities; Kristin Running, University of Minnesota - Twin Cities

This study described a stratified computerized adaptive test (CAT) administrations designed to provide cognitive diagnostic feedback to test taker regarding his or her weaknesses across subdomains. Through *R*-programmed simulation, the study evaluated the parameter recovery performances of both methods against regular Rasch based CAT and the diagnostic feedback accuracy.

Electronic Board #15

Two General Iterative Q-Matrix Validation Procedures

Ragip Terzi, The Turkish Ministry of National Education; Jimmy de la Torre, The University of Hong Kong

This study proposes two iterative indices, iJSD and iGDI, to determine the correctness of attribute specifications in the Q-matrix under a generalized CDM. The indices can identify misspecified q-entries at a high rate, specifically, when attributes are correlated, and the false-negative rate is around the nominal level under favorable conditions.

Electronic Board #16

Estimating Composite Score Reliability for Mixed-Format Tests: Three Theories

Jaime Malatesta, The University of Iowa; Kuo-Feng Chang, The University of Iowa; Mingqin Zhang, The University of Iowa; Won-Chan Lee, The University of Iowa

This study demonstrates different methods for computing composite score reliability for mixed-format tests that also contain testlets. Reliability indices based on classical test theory, generalizability theory, and item response theory are compared and contrasted using data from three Advanced Placement exams that vary with respect to content and test structure.

Electronic Board #17

A Cross-classified modeling approach to teacher-rating data in a Bayesian framework

Jayashri Srinivasan, University of California, Los Angeles; Michael Seltzer, University of California, Los Angeles

This study employs a cross-classified modeling approach to examine teacher-rating data with a modest number of raters in a Bayesian framework. Also, using data from a teacher portfolio study, we apply an IRT measurement model to further investigate the polytomous teacher-practice items, which are increasingly common in educational assessments.

Electronic Board #18

Modeling Item Position and Context Effects in Formative CAT

Anthony Albano, University of Nebraska-Lincoln; Liuhan Cai, University of Nebraska-Lincoln

In CAT, the position and context of an item may change from precalibration to operational administration, potentially biasing the estimation of item and person parameters. This study demonstrates procedures for modeling position and context effects using data from a formative assessment program. Implications for parameter invariance in CAT are discussed.

Electronic Board #19

Estimation of Polychoric Correlation and Structural Equation Model Parameters Under Nonnormality

Scott Monroe, UMass Amherst

In multistage estimation for ordinal data structural equation models, multivariate normality is typically assumed for the underlying response variables. This research explores how the type of nonnormality and threshold values may lead to substantial bias in both the polychoric estimates and the structural model parameter estimates.

Electronic Board #20

Expectation-Maximization-Maximization for 1PL-AG Model

Shaoyang Guo, Jiangxi Normal University; Tong Wu, University of Illinois at Urbana-Champaign; Chanjin Zheng, Jiangxi Normal University; Wen-Chung Wang, The Education University of Hong Kong

1PL-AG model with a modest sample size remains a challenge. This study proposes an Expectation-Maximization-Maximization (EMM) algorithm for this model. The results indicate that EMM has huge advantages in relative small samples (about 1000) over the tradition EM algorithm and can promote the widespread application of the model.

Saturday, April 14, 2018
12:25-1:55pm, Ambassador III, Coordinated Sessions, C1

The History of Educational Testing form 1950 to the Present

Session Chair: Brian Clauser, National Board of Medical Examiners

The purpose of this session is to provide an overview of how testing and psychometric theory have evolved since 1950. Robert Brennan will provide an integrated history of the development of generalizability theory and classical test theory beginning with the publication of Cronbach's landmark 1951 paper on coefficient alpha. Ronald Hambleton will describe the evolution of item response theory from Gulliksen's initial call for models that would lead to parameter invariance within classical test theory to the development of what has become known as modern test theory. Michael Kane and Brent Bridgman describe the evolution of validity theory from the introduction of construct validity through Messick's work to current conceptualizations of score validity as a structured argument. Michael Bunch will discuss how federal legislation has shaped educational measurement during that period.

History of Item Response Theory Models and Applications
Ronald Hambleton, University of Massachusetts, Amherst

Classical Test Theory and Generalizability Theory: An Integrated History (1950-Present)
Robert Brennan, University of Iowa

Saturday, April 14, 2018
12:25-1:55pm, Broadway I, Coordinated Sessions, C2

Addressing Motivational Issues in Low-Stakes Testing: U.S. and International Perspectives

Session Chair: Ou Liu, Educational Testing Service

Students' test-taking motivation in a low-stakes testing situation has long been investigated. Research to date suggests that students regularly exhibit low effort when taking low-stakes tests and that low motivation negatively impacts test performance. The need to study motivation becomes increasingly important with the prevalence of international assessments such TIMSS, PISA, and group-level assessments employed by higher education institutions to evaluate students' learning. The four presentations here will shed new light on this important topic by including both classroom-based and large-scale assessments and examining both U.S. and international samples. Presentation 1 discusses an experimental study designed to motivate Chinese students to take a college-level critical thinking assessment. Presentation 2 reports on how students' test-taking motivation, indicated by their item response time, affects the comparison of aggregated PISA scores. Presentation 3 explores how Sweden students' test-taking motivation, indicated by students' responses to a test-taking motivation scale, varies across age groups and in relationship to TIMSS performance. The last presentation examines whether test experience has an impact on students' self-report efforts. The four presentations jointly advance our understanding of the impact of motivation on test scores, by subgroups and over time, and strategies that can be used to enhance motivation.

The Motivational Effect on a Critical Thinking Test: A Chinese Study
Ou Liu, Educational Testing Service; Joseph Rios, Educational Testing Service; Guangming Ling, Educational Testing Service; Liping Ma, Beijing University

The Impact of Differential Test Taker Engagement on Aggregated Scores
Steve Wise, Northwest Education Association; James Soland, Northwest Education Association; Yuanchao Bo, Northwest Education Association

Test-Taking Motivation in Swedish TIMSS/TIMSS Advanced: Findings Across Cohorts and Over Time
Hanna Eklöf, Umeå University; Denise Costa, National Institute for Educational Studies

Analysis of Change: Examining the Effect of Attributional Bias on Self-reported Motivation
Aaron Myers, James Madison University; Sara Finney, James Madison University

Saturday, April 14, 2018
12:25-1:55pm, Broadway II, Coordinated Sessions, C3

Measuring Clinical Judgment in Nursing: Integrating Technology Enhanced Items

Session Chair: Joe Betts, Pearson Vue

This coordinated session will highlight the recent research of a large-scale testing programs efforts to expand the exam to include the complex construct of clinical judgment (CJ). The session will describe the evolution of this research process. The papers proceed in a chronological fashion from the first paper that discusses the unique job task analysis that uncovered the emerging need, and the advancement of the R&D process that merged a cognitive psychological decision-making model with the daily praxis outlined by the job task analysis to a definable assessment model will be explored. The second presentation will highlight the process by which new technology enhanced item (TEI) types were identified for their potential utility in measuring aspects of model along with the methods used to validate design issues. The third paper will outline a number of possible scoring rubrics that were designed in an attempt to align scoring more closely to the proposed underlying cognitive construct. The final paper will then explore both a signal-detection and polytomous item response theory (IRT) framework for evaluating items and unique methods for providing feedback to item developers related to new item types.

Moving a Traditional Assessment into the Next Generation: Exploring the Road Ahead
Ada Woo, NCSBN; Doyoung Kim, NCSBN; Joe Betts, Pearson VUE; William Muntean, Pearson VUE; Xiao Luo, NCSBN

Defining New Item Types for a Clinical Judgment Construct
Joe Betts, Pearson VUE; Ada Woo, NCSBN; William Muntean, Pearson VUE; Doyoung Kim, NCSBN

Evaluating Scoring Models to Align with Proposed Cognitive Constructs Underlying Item Content
Doyoung Kim, NCSBN; Ada Woo, NCSBN; Joe Betts, Pearson VUE; William Muntean, Pearson VUE

Using Signal-detection Theory to Enhance IRT Methods: A Clinical Judgment Example
William Muntean, Pearson VUE; Joe Betts, Pearson VUE; Xiao Luo, NCSBN; Doyoung Kim, NCSBN; Ada Woo, NCSBN

Saturday, April 14, 2018
12:25-1:55pm, Broadway III, Coordinated Sessions, C4

Classroom Assessment and Educational Measurement

Session Chair: Susan Brookhart, Duquesne University
Session Discussant: James McMillan, Virginia Commonwealth University

Recently, NCME has recognized and been interested in better understanding how classroom assessment perspectives can inform educational measurement and how educational measurement perspectives can inform classroom assessment. Accordingly, the purpose of this coordinated session will be to highlight work in the area of classroom assessment and educational measurement. As a group, these presentations describe growing edges of scholarship and demonstrate that the two fields (classroom assessment and educational measurement) can work together to deepen understandings in both fields. The result is enhanced understanding of evidence of what students know and can do in classroom settings as well as expanded understanding of measurement theory and issues.

Cognitive Skill Diagnosis is Insufficient: The Challenge Measuring Learning with Classroom Assessments
Jacqueline Leighton, University of Alberta

Language as Mediator of Valid Interpretations of Information Generated by Classroom Assessment
Alison Bailey, University of California, Los Anglese; Richard Duran, University of California, Santa Barbara

Guidance in the Standards for Classroom Assessment Practices to Support Instructional Decisions
Steve Ferrara, Measured Progress; Kristen Maxey-Moore, Denver Public Schools; Susan Brookhart, Duquesne University

Supporting Students to Notice, Interpret, and Act on Evidence of Learning
Caroline Wylie, Educational Testing Service; Christine Lyon, Educational Testing Service

Saturday, April 14, 2018
12:25 – 1:55, Gershwin 2, Invited Session, C5

Creating the Capacity to Increase Understanding of What Works in Schools, How It's Measured and Why It Works

Session Chair: Dr. Geoffrey Maruyama, National Association of Assessment Directors, and University of Minnesota

The most immediately pressing issues facing U.S. public schools today are national in scope and relatively limited in number. They tend to be tied to a fundamental purpose of education, namely, creating an educated workforce for a 21st century participatory democracy. Insofar as the biggest issues facing schools tend to be common across schools, many schools (through their assessment and research departments) likely are attempting to measure and assess effectiveness of similar if not identical approaches to address challenges. Many others not actively doing evaluation or research are collecting outcome measures that could be used for research on effectiveness of the programs they are doing. The National Association of Assessment Directors (NAAD) have the capacity to share data and knowledge across districts about what things work, under what conditions, how they are measured and why they work. Therefore, we propose that NAAD begin a process that will identify areas of greatest common interest and potential impact, and determine how we can increase sharing of findings and perhaps de-identified data across districts to scale up the level of ongoing work. This process begins with this symposium.

Building a Sustainable, Collaborative Approach to Provide Information That Drives Broader Adoption of Effective Practices in Schools
Geoff Maruyama, University of Minnesota

Using Assessment Practices in Service of Equitable Outcomes
Adrienne Bailey, Panasonic Foundation; Samuel Etienne, Elizabeth New Jersey School District

Defining Redefining Ready: Common Measures to Determine College Readiness Beyond Test Scores
Jeffery Smith, Township High School District 214, Illinois

STEAM - Is it Art or Creative Thinking: What Do We Measure and How Do We Measure it?
Bonnie Strykowski, Mesa Public Schools, Arizona

Bringing to Life District Assessment Systems: Aligning Local and State Assessments with Funding for Student Success and School Improvement
Antoinette Stroter, Winston-Salem/Forsyth County Schools, North Carolina

Saturday, April 14, 2018
12:25-1:55pm, Belasco, Individual Presentations, C6

Something's Missing: Working with Incomplete Data

Session Discussant: John Donoghue, Educational Testing Service

Using Nonresponse Times to Account for Omitted Items in Competence Tests
Esther Ulitzsch, Freie Universität Berlin; Steffi Pohl, Freie Universität Berlin; Matthias von Davier, National Board of Medical Examiners

A hierarchical framework for the joint modeling of response and nonresponse behavior is presented, which utilizes the information provided by nonresponse times to model mechanisms underlying item omissions. The advantages of the proposed framework are illustrated using both a simulation study and an empirical data analysis.

Imputation Methods to Deal with Missing Responses in Computerized Adaptive Multistage Testing
Dee Duygu Cetin-Berber, University of Florida; Halil Ibrahim Sari, Kilis 7 Aralik University

The purpose of this study is to investigate multiple missing data handling approaches in multistage testing (MST). Performance of four missing data imputation techniques is examined against treating missing items as incorrect responses. It is hypothesized that imputation methods will perform better compared to treating missing items as incorrect responses.

Using repeated ratings to improve measurement precision in incomplete rating designs.
Eli Jones, University of Missouri; Stefanie Wind, University of Alambama

This simulation study explores the effects of including multiple ratings per examinee per judge on the on the precision of examinee estimates in rater-mediated assessments with sparse rating designs. Using a Rasch model, we found that repeated ratings increased measurement precision and parameter recovery levels for examinees.

Explaining Omission Tendency in a Mathematics Test for German SEN Students
Nicole Haag, Institute for Educational Quality Improvement

Students with special educational needs (SEN) tend to omit test items in large-scale assessments. In a nationwide German mathematics assessment for ninth graders, item type and students' test-taking motivation predicted missing values in several ability-tracked groups of students. However, omitting linguistically complex items was particularly likely for SEN students.

The Impact of Missing Contextual Questionnaire Data on Plausible Value Generation
Xiaying Zheng, University of Maryland, College Park; Young Yee Kim, American Institutes for Research; Lauren Harrell, National Center for Education Statistics; Markus Broer, American Institutes for Research

In large-scale assessments, plausible values (PVs) are generated for secondary analysts by conditioning on demographic and contextual questionnaire (CQ) variables. This research investigates the implications of a state opting out of CQ in 2015 National Assessment of Educational Progress. The impacts on PVs for national- and subgroup-level inferences are examined.

Saturday, April 14, 2018
12:25-1:55pm, Plymouth, Individual Presentations, C7

Moving forward with MST

Session Discussant: Duanli Yan, Educational Testing Service

Multistage Testing with Dual Purposes: Ability Estimation and Classification
Yanming Jiang, Educational Testing Service

We focus on a multistage test (MST) design for an achievement assessment that has a classification purpose. Automated test assembly with additional constraints on the classification threshold is proposed. The accuracy of both ability estimation and classification is examined for both the new test assembly method and three routing methods.

Optimizing the Design of a Multistage Adaptive Test
Yuehmei Chien, Pearson; Hui Deng, The College Board

Given the numerous possible variations in the design of multistage adaptive tests (MST), this simulation study explores an optimal set of the MST design features for a large-scale assessment when the item pool is limited and a large number of content constraints are employed in test construction.

Impact of item pool characteristics on MST form assembly
Xuechun Zhou, NCS Pearson; Qi Diao, ACT; Liyang Mao, IXL Learning

The purpose of this study is to provide a method for designing and revising item pool blueprint to optimize MST module assembly. The optimal item pool is determined using the *p*-optimality method. Form assembly using the optimal and operational item pools is evaluated by the pool utilization and MST simulations.

Alternative Multistage Adaptive Testing Designs for Items Requiring Item-level Accommodation
EunHee Keum, UCLA/CRESST; Hansen Mark, UCLA/CRESST

This study evaluates the performance of alternative multistage testing designs developed to administer items requiring item-level accommodation. Results obtained with these designs are compared against those from a fully adaptive test and the fixed forms with respect to score precision and reliability via simulation study.

Routing Strategies for Multidimensional Multistage Tests
Hyung Jin Kim, The University of Iowa

As multistage testing starts to measure multiple traits using multidimensional item response theory models, it is crucial that routing for multidimensional multistage testing is conducted as successfully as possible. This study investigates, for multidimensional datasets, how accurately examinees are classified at the last stage for different routing rules.

Saturday, April 14, 2018
12:25-1:55pm, Manhattan, Individual Presentations, C8

Designing and Evaluating Tests With and Without IRT

Session Discussant: Alison Ames, James Madison University

Simulating the Complex, Dynamic Nature of Teaching and Learning through Systems Dynamics
Pamela Paek, ACT, Inc.; Britte Cheng, SRI International

Using systems dynamics, we mathematically model the complex and dynamic nature of assessment practices, including interactions and relationships that are currently not modeled or measured. We describe a runnable simulation we created, integrating multiple types of student outcomes that provide new perspectives on how data can be evaluated and triangulated.

Comparison of vertical scaling methods for measuring Spanish early literacy growth
Patrick Meyer, University of Virginia; Karen Ford, University of Virginia; Marcia Invernizzi, University of Virginia

PALS español involves tests of early literacy in Spanish administered and scored by teachers at four time points. Vertical scaling with multidimensional and unidimensional IRT models indicated the testlet model was best, but there was no detrimental effect of using a Rasch model, which is more suited to teacher scoring.

Item Response Theory Models for Ipsative Tests with Multidimensional Partial Ranking Items
Xue-Lan Qiu, Department of Psychology, The Education University of Hong Kong; Wen-Chung Wang, Department of Psychology, The Education University of Hong Kong; Chia-Wen Chen, Assessment Research Centre, The Education University of Hong Kong; Sage Ro, IBM

Developments of IRT models for ipsative tests with multidimensional forced-choice items have been witnessed in recent years. In this study, we introduced a new IRT model for multidimensional partial ranking items, provided an empirical example and conducted a brief simulation study to evaluate the parameter recovery of the new model.

Explanatory Item Response Modeling of an Algebra Concept Inventory
Claire Wladis, CUNY Graduate Center; Jay Verkuilen, CUNY Graduate Center; Sydne McCluskey, CUNY Graduate Center

Algebra is particularly important due to its status as a ``gatekeeper''. We use explanatory item response modeling (EIRM) to examine differential item/test function on an elementary algebra concept inventory. The sample was gathered at a large Northeastern community college. We have coded item features and student characteristics.

Saturday, April 14, 2018
12:25-1:55pm, Ambassador II, Individual Presentations, C9

Reflecting on Item and Form Development

Session Discussant: Laurie Davis, ACT

Development of interactive work simulation item templates using the Assessment Engineering Framework
Ian Clifford, Prometric; Aolin Xie, Prometric

Interactive simulation item templates were developed using Assessment Engineering (AE). AE incorporates psychometric and content targets into the design and utilizes automatic item generation. We present methodology through the AE framework from cognitive and evidence model through to production from task and item templates followed by summary of empirical results.

A simulation-based method for the optimal number of options for multiple-choice items
Hongwen Guo, Educational Testing Service; Jiyun Zu, Educational Testing Service; Patrick Kyllonen, Educational Testing Service

We proposed a simulation-based method to evaluate effects of different numbers of options on test characteristics. Grounded on theory and literature, we use two criteria (low frequency and poor discrimination) to remove nonfunctioning options and two schemes (random and educated guessing) to model hypothetical response behavior for the removed options.

The Expanded Evidence-Centered-Design (e-ECD) Framework for Learning & Assessment Systems
Meirav Arieli-Attali, ACT; Alina von Davier, ACT; Benjamin Deonovic, ACT

This study presents an expansion to the ECD framework by incorporating aspects of learning into each of the three core models of ECD, as well as integrating methodology from computational psychometrics. The new framework and its implementation in designing a learning & assessment system for science will be presented.

How to Determine What Item Screening Criteria to Use
Bozhidar Bashkov, American Board of Internal Medicine; Jerome Clauser, American Board of Internal Medicine

Successful testing programs rely on high-quality test items to produce defensible exams and reliable scores. However, what criteria do test items need to meet to be deemed psychometrically acceptable? This study demonstrates an empirical approach to determining the screening criteria for a given testing program and purpose.

Constructing Parallel Forms for Generalized Partial Credit Model: An Item-matching approach
Pei-Hua Chen, National Chiao Tung University; Cheng-Yi Huang, National Chiao Tung University

The purpose of this study is to apply item-matching approach for multiple parallel forms assembly based on the Generalized Partial Credit model. A real item bank consists of 210 items was used. The results showed that item-matching approach can produce similar results as integer linear programming approach.

Saturday, April 14, 2018
12:25-1:55pm, Gershwin I, Graduate Student Research Session, C10

GSIC Graduate Student Poster Session 1

Electronic Board #1

Multiple-Cycle Expectation Maximization Item Response Theory Scale Linking with Mixed-Format Tests

Alex Brodersen, University of Notre Dame; Ian Campbell, University of Notre Dame; Ying Cheng, University of Notre Dame

Linking in important process in test development using item response theory. Multiple-Cycle Expectation Maximization (MEM) is a fixed parameter linking method recently extended to polytomous item types (Zheng, 2016). The current paper extends MEM to linking in mixed-format tests and evaluates the technique via simulation study.

Electronic Board #2

Examining the Effect of Best Distractor Location on Item Difficulty

Jinnie Shin, University of Alberta; Okan Bulut, University of Alberta; Mark Gierl, University of Alberta

This empirical study investigated whether the position of the best (i.e., most attractive) distractor had any impact on the difficulty level of multiple-choice items. The results indicated that the best distractor position and the distance between the best distractor and the correct response option significantly affected item difficulty.

Electronic Board #3

Comparison of Methods for Treatment of Differential Item Functioning

Xiaowen Liu, University of Connecticut, Neag School of Education; H. Jane Rogers, University of Connecticut, Neag School of Education

The current study compared four treatments for items that have been identified as showing DIF: deleting; ignoring, multigroup modeling, and modeling DIF as a secondary dimension. Results of a simulation show that a multigroup modeling approach for DIF items produces the most accurate and least biased trait estimates.

Electronic Board #4

Evaluating Competing MIRT Models Using Different Goodness of Fit Statistics

Xinchu Zhao, University of South Carolina; Brian Habing, University of South Carolina

The purpose of this study is to evaluate the effectiveness of different goodness of fit statistics at distinguishing between the compensatory (CM), non-compensatory (NCM), and rotatable variable compensation (RAVCM) MIRT models. The fit indices investigated include log likelihood, AIC, BIC, and DIC.

Electronic Board #5

A Comparison of IRT Scoring Methods for Mixed-Format Multistage Tests

Shumin Jing, University of Iowa; Kyung Yong Kim, University of North Carolina at Greensboro; Won-Chan Lee, University of Iowa

The purpose of this study is to evaluate the performance of various IRT scoring methods for mixed-format multistage tests. Expected A Posteriori estimation produced the most accurate results compared to the Maximum Likelihood estimation, Bayesian model estimation, and Weighted Likelihood estimation.

Electronic Board #6

Assessing Classification Equity Property in IRT Equating

Mingqin Zhang, The University of Iowa; Seohee Park, The University of Iowa; Kyung Yong Kim, The University of North Carolina at Greensboro

A simulation study was conducted to investigate the classification equity property in IRT observed- and true-score equatings with a common-item nonequivalent groups design. IRT-recursive-based classification consistency and accuracy indices were used to evaluate reliability and validity of classification decisions on equated forms.

Electronic Board #7

Factors affecting the utility of Rasch Trees for detecting Differential Item Functioning

Elizabeth Patton, University of North Carolina Greensboro

Rasch Trees pose a new methodology for DIF detection. However, more research is needed to investigate the methodology's power, Type I error, and accuracy. This research focused on the impact of sample size, covariate correlation and type. Decreasing the sample size and increasing the correlation caused high levels of inaccuracy.

Electronic Board #8

Rater effects on equating mixed-format test in common-item nonequivalent groups design

Yoon Ah Song, University of Iowa

This simulation study is to explore to what extent different kinds of rater effect (no effect, leniency, and severity) affect on the results of equating mixed-format test in common -item nonequivalent design. The results will be discussed in terms of the reliability among various equating methods.

Electronic Board #9

Model Selection for IRT Observed-Score Equating for Mixed-Format Tests that Contain Testlets

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

The choice of item response theory (IRT) models on equating for testlet-based tests has been widely discussed. However, related literature has generally been confined to tests composed of a single item type. Therefore, this study aims to examine the impact of IRT models on equating for mixed-format tests containing testlets.

Electronic Board #10

Bayesian, Method of Moments, and REML Estimation in Mixed-Effects Reliability Generalizations Studies

Brandie Semma, Texas A&M University; Maria Henri, Texas A&M University; Wen Luo, Texas A&M Unviersity

Reliability generalization is a meta-analytic technique assessing variability in reliability estimates across studies. However, many disagreements across several methodological issues exist. The purpose of this study is to compare the performance of various estimation methods in RG studies through a simulation and an illustrative example.

Electronic Board #11

Comparison between DINA model and Noncompensatory MIRT model

Mingqi Hu, University of Illinois at Urbana-Champaign

Research compares noncompensatory MIRT with DINA model, to evaluate its cognitive diagnostic function. In simulation, CP method transforms continuous latent traits into categorical variables. N = 2000. 2 (attributes = 3 or 6) by 3 (item lengths =15, 30 or 60) conditions. PCCR and ACCR are calculated for estimation accuracy.

Electronic Board #12

Looking for a Consensus about the Concept of Validity: A Delphi Study

Sandra Camargo, Universidad Nacional de Colombia; Aura Herrera, Universidad Nacional de Colombia; Anne Traynor, Purdue University

We aimed to identify, using a Delphi study, precise aspects of the concept of validity about which there is, and is not, current consensus. Study participants included academic experts who have led the discussion on the concept of validity in recent decades.

Electronic Board #13

Modeling Response Styles in Cross-Country Self-Reports Using a Multilevel-Multidimensional Nominal Response Model

Unhee Ju, Michigan State University; Carl Falk, McGill University

This study examines the effects of extreme response style (ERS) on self-rated scores of teachers' self-efficacy across countries using a multilevel-multidimensional nominal response model. Utility of the model regarding across-country comparisons, changes in the relationships among latent traits, and proneness of particular items to ERS are illustrated.

Electronic Board #14

Elimination scoring versus correction for guessing: A simulation study

Qian Wu, KU Leuven; Tinne De Laet, KU Leuven; Rianne Janssen, KU Leuven

This study simulates the impact of correction for guessing versus elimination scoring on how examinees answer multiple-choice questions. A two-step approach is used to predict answering patterns, combining a psychometric model accounting for ability with decision theory accounting for individual differences in risk aversion.

Electronic Board #15

Testing Dynamic Complementarity in Educational Opportunities to Accumulate Relevant Human Capital

Gulsah Gurkan, Boston College; Diego Luna Bazaldua, National Autonomous University of Mexico; Henry Braun, Boston College

We report on a test of Heckman's "Dynamic Complementarity" hypothesis: Concatenating effective interventions yields multiplier effects. We employed longitudinal data from a school based intervention that addresses non-academic barriers to learning. We find no evidence of Dynamic Complementarity on test scores, but uncover a methodological paradox that merits further study.

Electronic Board #16

A Nonparametric Computerized Adaptive Testing for Cognitive Diagnosis in Classroom

Yuan-Pei Chang, National Taiwan Normal University; Chia-Yi Chiu, Rutgers, The State University of New Jersey; Rung-Ching Tsai, National Taiwan Normal University

An innovative computerized adaptive testing for cognitive diagnosis based on the nonparametric classification method (Chiu & Douglas, 2013) is proposed in the study. The proposed item selection method does not rely on any item parameter calibration and thus can be used to analyze samples of all sizes.

Electronic Board #17

Comparing Item Exposure Rate and Test Security for Computerized Single -and Multiple-Pools

Qiao Lin, University of Illinois at Chicago; Haiqin Chen, American Dental Association

The purpose of this study is to examine the item exposure rates and test security in computerized adaptive testing. A simulation study is conducted to investigate: 1) how the number of pools affects the exposure rate; 2) benefits of multiple pools when a certain number of items are compromised.

Electronic Board #18

A Psychometric Analysis of the NU Data Knowledge Scale

Pamela Trantham, University of Nebraska - Lincoln; Jonathon Sikorski, Munroe-Meyer Institute for Genetics and Rehabilitation.; Rafael de Ayala, University of Nebraska - Lincoln; Beth Doll, University of Nebraska - Lincoln

The NU Data Knowledge Scale is a measure of teacher data literacy. The psychometric properties were examined, finding it to be a reliable, unidimensional measure. The use of an IRT model was investigated. The 1PL model provided the best fit. A concordance table was created to quickly ascertain teacher ability.

Electronic Board #19

Evaluating the Dimensionality of Multistage-Adaptive Test Data

Maritza Casas, University of Massachusetts Amherst

Assessing dimensionality in multistage adaptive tests is challenging due to the sparseness of the item response matrix. The purpose of this study is to evaluate the performance of the Full Information Maximum Likelihood (FIML) procedure to estimate the dimensionality of a multistage-adaptive test, which by design involves non-random missing data.

Electronic Board #20

Psychometric Results for Multiple Methods of Scoring Scales Measuring Social Desirability

Murat Kilinc, University of Iowa; Walter Vispoel, University of Iowa; Carrie Morris, University of Iowa

We compared fifteen methods for scoring measures of social desirability. Seven-point scoring yielded stronger evidence of reliability and concurrent validity, whereas new methods with fewer options emphasizing exaggerated responses produced fewer false-positive errors in flagging faking. Results underscore that decision makers should cater scoring methods to particular uses of scores.

Electronic Board #21

Evaluating item fit in the presence of learning

Ben Stenhaug, Stanford; Ben Domingue, Stanford

This paper examines how dynamic abilities, likely to be an issue in learning environments such as MOOCs, affect evaluations of item fit. Interpretation of fit statistics is shown to be complicated by both differential rates of growth as well as different patterns of growth.

Electronic Board #22

The Relationships among Motivation, Prior-Knowledge, Engagement, and Achievement in a MOOC

Jingxuan Liu, Georgia State University; Hongli Li, Georgia State University

Increasing student engagement level is challenging and important for MOOC design, retention, and completion. It is valuable to identify factors influencing student MOOC engagement and achievement. However, such literature remains thin. The present study examines relationships among student motivation, prior-knowledge, engagement, and achievement in MOOCs.

Electronic Board #23

IRT Observed Score Equating Using MCMC

Seohee Park, University of Iowa; Kyung Yong Kim, University of North Carolina at Greensboro

This study demonstrates that the MCMC method can be used to simultaneously estimate equating relationships, uncertainty of equating, and intervals. Additionally, equating relationships and uncertainty of equating obtained with MCMC estimates are compared with equating relationships and random equating error obtained with MMLE estimates and the bootstrap method, respectively.

Saturday, April 14, 2018
2:15-3:45pm, Ambassador III, Coordinated Sessions, D1

The History of Educational Testing form 1950 to the Present

Session Chair: Brian Clasuser, National Board of Medical Examiners

The purpose of this session is to provide an overview of how testing and psychometric theory have evolved since 1950. Robert Brennan will provide an integrated history of the development of generalizability theory and classical test theory beginning with the publication of Cronbach's landmark 1951 paper on coefficient alpha. Ronald Hambleton will describe the evolution of item response theory from Gulliksen's initial call for models that would lead to parameter invariance within classical test theory to the development of what has become known as modern test theory. Michael Kane and Brent Bridgman describe the evolution of validity theory from the introduction of construct validity through Messick's work to current conceptualizations of score validity as a structured argument. Michael Bunch will discuss how federal legislation has shaped educational measurement during that period.

An Overview of Trends in Validity Theory, 1950 to the Present
Michael Kane, Educational Testing Service; Brent Bridgeman, Educational Testing Service

The Federal Role in Shaping Educational Measurement Practice: 1950-Present
Michael Bunch, Measurement Incorporated

Saturday, April 14, 2018
2:15-3:45pm, Broadway I, Coordinated Sessions, D2

Using Classification-based Psychometrics in Local Assessment Systems for Feedback and Accountability

Session Chair: Laine Bradshaw, University of Georgia
Moderator: Susan Weigert, United States Department of Education
Panelist: Laine Bradshaw, University of Georgia
Panelist: Hua-Hua Chang, University of Illinois-Urbana Champagne
Panelist: Scott Marion, NationalCenter Improvement of Educational Assessment
Panelist: Jonathan Templin, University of Kansas

Diagnostic classification modeling (DCM) has been shown to be a statistically well-established methodology for classifying students according to mastery levels of multiple attributes. Because they are efficient methods whose feedback provides meaningful groupings of students, DCMs are well-suited to meet educators' needs for ongoing, detailed feedback about what students know well and what they need additional help to learn. DCMs also provide a different perspective on accountability, a perspective where progress is not assumed to exist on one continuum. While much research has been conducted on the theoretical properties of the models, few efforts have applied the methods in operational testing programs.

This panel will discuss the opportunities and challenges surrounding efforts to transition innovative DCM methodology to our classrooms. The panelists will include three professors, each of whom are closely partnering with school systems to create and implement a DCM-based assessment system, and the Executive Director of National Center for the Improvement of Educational Assessment who has guided states in the design and implementation of innovative assessment systems. The facilitator of the panel is a member of the National Initiatives Team from the US Department of Education whose work includes guiding innovative assessment initiatives.

Saturday, April 14, 2018
2:15-3:45pm, Broadway II, Coordinated Sessions, D3

Response times in educational measurement: Moving beyond the simple structure hierarchical model

Session Chair: Jesper Tijmstra, Tilburg University

With the advance of computerized tests, recording response times in addition to response accuracy has become commonplace in educational measurement. In the last decade the hierarchical modeling framework for response times and accuracy (van der Linden, 2007) has become the standard approach for jointly modeling response time and accuracy in educational measurement. Here, separate measurement models are considered for accuracy and time, which are connected on the higher level by considering the relationships between the parameters in the two parts of the model. Although it succeeds in providing a clear structure for studying both response times and accuracy, this modelling framework is based on a set of assumptions which may not match the complex picture that arises when realistic response processes are considered. In this symposium, these assumptions - primarily the simple structure assumption and the assumption of conditional independence between response times and accuracy - will be considered critically, and statistical models that relax these assumptions will be proposed.

Improving precision of ability estimation: Getting more from response times
Jesper Tijmstra, Tilburg University

Hidden Markov Mixture Modeling of Responses and Categorized Response Times
Dylan Molenaar, University of Amsterdam

Response moderation models for conditional dependence between response time and accuracy
Maria Bolsinova, University of Amsterdam

The meaning of residual dependencies between response time and response accuracy
Paul De Boeck, Ohio State University

Saturday, April 14, 2018
2:15-3:45pm, Broadway III, Coordinated Sessions, D4

Measuring Collaboration and Engagement using "Big Data"

Session Chair: Maria Bertling, Harvard University
Session Chair: Andrew Ho, Harvard University
Session Discussant: Arthur Graesser, University of Memphis

There is growing evidence that motivational and collaborative factors predict long-term individual outcomes (Almlund et al., 2011; Deming, 2015). The ability to actively engage in solving meaningful problems while collaborating with others is further emphasized in 21st century skills frameworks that outline constructs critical for success in modern society (Trilling & Fadel, 2009). However, many practical and theoretical challenges remain in measuring these constructs. The majority of instruments are based on self-reported data, which have limited reliability and poor properties for measuring change over time (Duckworth & Yeager, 2015).

As digital measurements proliferate in interactive assessment systems, there are new possibilities for assessment of observed behaviors that indicate engagement and collaboration. This symposium includes four papers that use digital "big data" improve measurement of collaboration and engagement. The first two papers employ computational psychometrics and multimodal learning analytics, respectively, to measure collaborative learning processes. The second two papers employ machine learning and latency response modeling, respectively, to measure engagement in online learning systems. Each presentation further identifies practical tips for implementing these measurements, as well as potential threats to generalizability.

New Directions in Assessing Collaborative Problem Solving Skills
Alina von Davier, ACT; Yigal Rosen, Harvard Univiersity; Kristin Stoeffler, ACT; Pravin Chopade, ACTNext

Capturing Collaborative Learning Processes Through Multi-Modal Sensing
Bertrand Schneider, Harvard University

The Faces of Engagement: Automatic Recognition of Student Engagement from Facial Expressions
Jacob Whitehill, Worcester Polytechnic Institute; Zewelanji Serpell, Virginia Commonwealth University

Measures of engagement in Massive Open Online Courses
Maria Bertling, Harvard University; Isaac Chuang, MIT

Saturday, April 14, 2018
2:15-3:45pm, Gershwin I, Coordinated Sessions, D5

Insight and Action: Diverse Perspectives on Critical Fairness Issues in Testing

Session Chair: Jessica Jonson, Buros Center for Testing - UNL
Session Discussant: Gregory Camilli, Rutgers University

In October 2017, the Buros Center for Testing hosted a diverse group of scholars from different professional fields in psychology to discuss gaps and new directions for research and practice in the areas of fairness in testing particularly in light of the 2014 *Standards for Educational and Psychological Testing*. This interactive gathering was made possible by AERA Research Conference funding. Key insights and recommendations from meeting participants will be shared in this coordinated session. This will include a summary of meeting results from the conference organizer along with detailed perspectives about fairness issues in testing and key meeting outcomes from scholars in educational measurement, school psychology, counseling psychology, and industrial/organizational psychology. The meeting activities were organized around three themes: methodological issues in measurement bias of scores, barriers in the opportunity to show true standing on a construct, and threats to validity of score interpretations for intended uses. An intended outcome for this coordinated session is to provide attendees a broader conception of cultural, language, and disability fairness issues in fields where testing is central and enhance thinking about what type of future methodological and applied research is needed to realize the practical aspirations of the *Standards*.

A School Psychologist's Perspective: Cognitive and academic evaluation of English learners.
Samuel Ortiz, St. John's University

A Counseling Psychologist's Perspective: Cultural considerations in psychological assessment
Lisa Suzuki, NYU Steinhardt

An Industrial/Organizational Psychologist's Perspective: Hiring a Diverse Workforce
Harold Goldstein, Baruch College - CUNY

An Educational Measurement Perspective: Estimating language-related measurement error when assessing English learners
Guillermo Solano-Flores, Stanford University

Saturday, April 14, 2018
2:15-3:45pm, Belasco, Individual Presentations, D6

Automatic Item Generation

Session Discussant: Hollis Lai, UAlberta

Human Machine Interactive Automatic Item Generation
Xinxin Zhang, University of Alberta; Mark Gierl, University of alberta

This study develops a two-module approach to automatically create item model and automatically generate items from the created model. We describe and demonstrate this new approach using surgical education test items with the self-developed interactive software.

A Cost-Benefit Analysis of Automatic Item Generation
Audra Kosh, MetaMetrics, Inc.; Mary Ann Simpson, MetaMetrics, Inc.; Lisa Bickel, MetaMetrics, Inc.; Mark Kellogg, MetaMetrics, Inc.; Ellie Sanford-Moore, MetaMetrics, Inc.; Ian Hembry, MetaMetrics, Inc.; Heather Koons, MetaMetrics, Inc.

We estimated the number of items that would have to be produced before the upfront costs of automatic item generation outweigh traditional item writing costs in the context of K-12 mathematics items. We considered time demands of item developers involved in each step of both manual and automated item development.

Automatic Item Generation Unleashed: Evaluation of a Large-Scale Deployment of Item Models
Yigal Attali, Educational Testing Service

The purpose of this study was to evaluate the results of a large-scale deployment of automatic item generation in an adaptive testing context, with a large number of item models, and a very large number of randomly generated item instances.

Integrating AIG into the Monte Carlo LOFT Algorithm to Reduce Item Exposure
John Weiner, PSI Services LLC; Gregory Hurtz, PSI Services LLC

Linear on the fly testing (LOFT) generates unique forms for each test-taker, but item exposure across forms may be of concern. We present metrics for projecting expected item exposure and form overlap, and demonstrate the benefits of integrating automated item generation (AIG) with LOFT to expand and protect item pools.

Saturday, April 14, 2018
2:15-3:45pm, Plymouth, Individual Presentations, D7

Developing CDM

Session Discussant: André Rupp, Educational Testing Service

A Sequential Higher-Order Latent Structure Model for Hierarchical Attributes

Peida Zhan, Beijing Normal University; Hong Jiao, University of Maryland, College Park; Wenchao Ma, University of Alabama; Shuliang Ding, Jiangxi Normal University

A sequential higher-order latent structural model (LSM) for hierarchical attributes in cognitive diagnosis was proposed. Unlike the regular higher-order LSM, by taking the sequential process/model into account, the proposed LSM is able to contain different attribute hierarchies and simultaneously retains the advantages of the higher-order latent structure.

A Diagnostic Tree Model for Multiple-Strategy Polytomous Responses

Wenchao Ma, The University of Alabama

This study develops a diagnostic tree model (DTM) for polytomous response data from constructed response items where multiple strategies are recorded. An MMLE-EM algorithm is developed to estimate item parameters of the DTM. Both simulation study and real data analysis are carried out to examine the viability of the DTM.

The General Q-Matrix Refinement Method

Yan Sun, Rutgers University; Yanhong Bian, Rutgers University; Chia-Yi Chiu, Rutgers University

Q-matrix which delineates the relationship between attributes and items in cognitive diagnostic models is subject to misspecification due to fallible judgments of experts. In this study, a general Q-matrix refinement method is proposed and evaluated in a simulation study, and the results showed its capability of recovering the true Q-matrix.

Multidimensional Higher-Order Models for Skills Diagnosis: Descriptive and Explanatory Approaches

Yoon Soo Park, University of Illinois at Chicago; Young-Sun Lee, Teachers College, Columbia University

This study proposes three models for analyzing skills mastery when there are multiple subject areas assessed, measuring attributes across different multidimensional higher-order latent traits. Models are examined using simulations and real-world data for (1) skills mastery and (2) explanatory relationships, where estimated attributes or latent traits serve as predictors.

Fitting a Diagnostic Assessment to Standards-Defined Skills versus Expert-Defined Skills

Aileen Reid, University of North Carolina at Greensboro; Karen Hoeve, University of North Carolina at Greensboro; Robert Henson, University of North Carolina at Greensboro

The study modeled a diagnosis assessment utilizing skills defined by content standards and skills defined by experts. Results show that using skills defined by experts improved both fit and interpretability, and provided useful diagnostic information to guide and improve teacher instruction and score reporting.

Saturday, April 14, 2018
2:15-3:45pm, Manhattan, Individual Presentations, D8

Technical Considerations in Assessing DIF

Session Discussant: Seock-Ho Kim, UGA

Assessing Differential Item Functioning in Continuous Items: A Comparison Study
Hsiu-Yi Chao, National Chung Cheng University; Jyun-Hong Chen, National Sun Yat-sen University; Ching-Lin Shih, National Sun Yat-sen University

This study proposed the MH method for continuous items (MHC), the multiple linear regression method (MLR), and the continuous SIBTEST procedure (CSIB) for DIF assessment in continuous items. Results of simulation study showed that the MLR and MHC outperformed the other methods in assessing uniform and nonuniform DIF, respectively.

Evaluating the cluster approach of differential item pair functioning in DIF analysis
Daniel Schulze, Freie Universität Berlin; Steffi Pohl, Freie Universität Berlin; Eric Stets, Freie Universität Berlin

When comparing competencies between groups of test-takers, items do not always display measurement invariance (called differential item functioning, DIF). We evaluated an assumption-free approach to DIF analysis by means of clustering relative DIF of item pairs. It performed well in a simulation study and is illustrated in an empirical example.

Measurement Noninvariance in a Thurstonian IRT Model
HyeSun Lee, California State University Channel Islands; Weldon Smith, University of Nebraska-Lincoln

The current simulation study examined power and Type I error rates in the detection of measurement noninvariance (MNI) in a Thurstonian IRT model and the impact of MNI on score estimation. It was found that power and Type I error rates were low, and the impact of MNI was substantial.

Information Criteria in the Study of Group Differences in Trace Lines
Seock-Ho Kim, The University of Georgia; Allan Cohen, The University of Georgia

A review of various information criteria is presented for the detection of differential item functioning (DIF) under item response theory (IRT). An illustration as well as results with simulated data are presented and contrasted with other DIF detection methods. Use of information criteria for general IRT model selection is discussed.

A Theoretical Power Formula for Crossing SIBTEST
Zhushan Li, Boston College

A theoretical power formula for Crossing SIBTEST is derived. The formula provides a means for sample size calculations in planning DIF studies with Crossing SIBTEST. Factors influencing the power are discussed. The correctness of the power formula is confirmed by simulation studies.

Saturday, April 14, 2018
2:15-3:45pm, Ambassador II, Individual Presentations, D9

Exploring Growth: Methods and Applications

Session Discussant: Thanos Patelis, HUMRRO

Reading Growth for Middle School Students with Significant Cognitive Disabilities
Daniel Farley, University of Oregon; Joseph Stevens, University of Oregon

Modeling growth for students with significant cognitive disabilities is complicated by test scaling, group heterogeneity, small samples, missing data, and status-based assessment designs. This study addressed these issues by: (a) using a common scale, (b) modeling growth for students in multiple demographic and exceptionality categories, and (c) using multiple cohorts.

A Hierarchical IRT Model for Identifying Group-Level Aberrant Growth to Detect Cheating
Jennifer Brussow, University of Kansas; William Skorupski, University of Kansas; William Thompson, University of Kansas

As cheating on high-stakes tests continues to threaten the validity of score interpretations, approaches for detecting cheating proliferate. Most research focuses on individual scores, but recent events show group-level cheating is also occurring. The present IRT simulation study extends the Bayesian Hierarchical Linear Model (BHLM) for detecting group-level aberrance.

Examining the Effectiveness of Anchoring Vignettes Longitudinally
Jiyun Zu, Educational Testing Service; Hongwen Guo, Educational Testing Serivce; Patrick Kyllonen, Educational Testing Service

Likert scale ratings may be incomparable because respondents may use the scale differently. Anchoring vignettes is a technique shown to reduce this problem in cross-sectional international surveys. We propose that anchoring vignettes may also be useful longitudinally for U.S.-only sample. We test our hypothesis by analyzing a longitudinal dataset.

Modeling Mediators of Within-Subject Change by Linear Growth Modeling Framework
Yusuf Kara, Southern Methodist University; Akihito Kamata, Southern Methodist University

This study introduces a linear growth modeling (LGM) framework for modeling the mediators of within-subject change on outcome variables measured over time. Parameter estimates from the analyses of an empirical dataset demonstrated that LGM approach is promising and can be preferred over multilevel modeling (MLM) approach for practical use.

Growth in Reading Comprehension and its Relationship with Mathematics and Science Development
Anthony Fina, Iowa Testing Programs

This study summarizes the development of reading, mathematics, and science achievement, and their interrelationships, for Grades 6-11 for a population of students using a latent growth model with three parallel processes. The impact of demographic and school-level variables are examined. Implications of this research for classroom instruction are discussed.

Saturday, April 14, 2018
2:15-3:45pm, Gershwin I, Individual Presentations, D10

Electronic Board Session 2

Electronic Board #1

Evaluating Construct Comparability between Paper- and Digital-Based Assessments with MIRT Application

Young Yee Kim, American Institues for Research; Soo Lee, American Institues for Research; Jiao Yu, American Institues for Research

NAEP is in a transition to digitally-based assessments (DBAs) from a paper-based assessments (PBAs). The possible introduction of a nuisance construct due to digital familiarity raises concerns in maintaining trend. This study examines construct comparability between PBAs and the DBA for grade 8 mathematics to inform trend decision in NAEP.

Electronic Board #2

A Further Investigation of the Generalized Dimensionality Discrepancy Measure for (Multi)Dimensionality Assessment

Shenghai Dai, Washington State Unviersity; Xiaolin Wang, The University of Kansas; Dubravka Svetina, Indiana University Bloomington

This study aims at (1) evaluating the performance of the (standardized) generalized dimensionality discrepancy measure ([S]GDDM) in assessing multidimensionality for models across factors including number of dimensions, test length, sample size, correlation between dimensions, and test structure complexity, and (2) providing conventions for the methods in determining the dimensional structure.

Electronic Board #3

Item-Level Predictive Validity on ACT Math and Science

James Gambrell, ACT, Inc.; Yu Su, ACT, Inc.

In the current study we analyze a longitudinal dataset to investigate relationships between item level characteristics on the ACT Math and Science tests and various college outcomes. Our analysis focused on predictive differences between items targeting different topics in math and science.

Electronic Board #4

Adaptive Scales and the Decision-Making Needed to Get There

Kimberly Colvin, University at Albany, SUNY; Michael Ellis, University at Albany, SUNY

With a practitioner in mind, this study documents the decisions to be made in the development of an adaptive scale. We start with piloting 233 Likert-type items, to scoring decisions, to deciding whether the psychometric properties of an adaptive scale better serves the researchers' needs than a linear scale.

Electronic Board #5

A Case of How Scaling Decisions Impact Psychometric Properties

Weiilng Deng, Educational Testing Service; Ourania Rotou, Educational Testing Service; Sandip Sinharay, Educational Testing Service; Neil Dorans, Educational Testing Service

Choice of and changes to a reporting scale have important implications for score interpretations, as well as consequences for test reliability and validity. This study shows scaling issues faced by real testing programs and how scaling decisions could lead to larger CSEM and negatively impact test reliability and validity.

Electronic Board #6

Language Comparability using Minimum Discriminant Information Adjustment

Hyeonjoo Oh, Educational Testing Service; Shameem Gaj, Educational Testing Service; Junhui Liu, Educational Testing Service

In this study, we investigated how the two language versions (i.e., English and Spanish) of the same test affect item level performance (e.g., item difficulty) and test level performance (e.g., test construct, scoring conversions) using the minimum discriminant information adjustment (Haberman, 1984).

Electronic Board #7

Evaluation of Dependence of Person Fit Statistics on Item Calibration Models

Wei Wang, Educational Testing Service; Sandip Sinharay, Educational Testing Service

The current study investigates and compares the dependence of three popular person fit statistics on item calibration models. The influences of various factors on the comparison results are also explored. Both simulated and real operational data are used.

Electronic Board #8

Evaluate the Effectiveness of Passage Exposure Control Mechanisms in Passage-Based CAT

Xin Li, ACT, Inc.; Meichu Fan, ACT, Inc.; YoungWoo Cho, ACT, Inc.

A series of simulations are carried out to implement and evaluate various item exposure control mechanisms on passages in passage-based computer adaptive testing. The effectiveness of those mechanisms are evaluated and compared under different pool quality conditions, along with the evaluation on estimation precision, pool utilization, and test overlap rate.

Electronic Board #9

Ethical Violations and Barriers to Good Practices in Psychological Testing in Turkey

Bengü Börkan, Boğaziçi University; Şeyda Çetintaş, boğaziçi university; Osman Yılmaz, Boğaziçi University; Gizem Öztemur, Boğaziçi University; Betül Gülcan, Boğaziçi Universitesi; Merve Özcan, Boğaziçi Universitesi

The experiences of test users working in Guidance and Research Centers in Turkey with the framework of International Test Commision's guidelines. In-depth interviews with 20 psychological show that test users practices can be addressed under the headings of insufficient physical environments, inadequate training and failing to adopt ethical principles.

Electronic Board #10

An Evaluation of Test Overlap in CAT Pools

Jie Li, ACT, Inc.; Yi He, ACT, Inc.; Chunxin Wang, ACT, Inc.

This study examined overall and conditional test overlap for a computer adaptive test (CAT). The relationship between test overlap, pool sizes, ability distributions, measurement precision and item exposure were evaluated. The test overlap results will provide supplementary information in CAT pool assembly and test security investigations.

Electronic Board #11

Comparisons of subscoring methods in computerized adaptive multistage testing

Jinah Choi, The University of Iowa

Subscores are of increasing interest due to their potential benefits of monitoring eximanee performance at subscale level. This research conducts a simulation study for comparing several methods for estimating subscores under various simulated adaptive multistage testing conditions. The results will provide useful guidelines relevant to practice.

Electronic Board #12

Design and Analyze the Computerized Adaptive Testing with the Graph Theory

Xiao Luo, National Council of State Boards of Nursing; Doyoung Kim, National Council of State Boards of Nursing

This study introduces a method of building and analyzing the graphical computerized adaptive testing (G-CAT) in order to visualize the internal process of CAT and conserve the testing efficiency in regular CAT. It gives test developers much greater controls over the test content and qualities before test administration.

Electronic Board #13

An Investigation of Parametric Bootstrap for S-X2/S-G2 Item Fit Measure

John Donoghue, Educational Testing Service; Adrienne Sgammato, Educational Testing Service

Assessing item-level IRT model/data fit of IRT is an ongoing challenge. Orlando & Thissen's (2001) goodness of fit measures, *S-X2* and *S-G2*, have shown promise. Findings of inflated Type I raise questions about using chi-squared as a reference distribution. This study examines the alternative of parametric bootstrap to assess significance.

Electronic Board #14

Increasing Underrepresented Minority Representation in Educational Measurement: An Analysis of Program Characteristics

Joseph Rios, Educational Testing Service; Jennifer Randall, University of Massachusetts Amherst; Marina Donnelly, University of Massachusetts Amherst

This study evaluated the availability of program-specific information that prospective ethnic minority students may find important in their application decision-making process. Findings suggest that educational measurement programs can work to improve greater flexibility in course and degree offerings as well as the types of information that they provide to applicants.

Electronic Board #15

Background Information and School Clustering Effects on Students' Opportunity-to-learn in PISA 2012

Diah Wihardini, Bina Nusantara University; Mark Wilson, UC Berkeley

Our study investigates the associations of background information with the newly-proposed opportunity-to-learn measures based on PISA 2012, after accounting for the school differences. Using multidimensional multilevel partial credit model with latent regression on Indonesian data, we presents how results can be utilized to leverage policy decisions for national education reforms.

Electronic Board #16

A Meta-Analytic Path Analysis of Academic Performance and Persistence

Paul Westrick, ACT; Huy Le, University of Texas, San Antonio; Steve Robbins, Educational Testing Service; Justine Radunzel, ACT; Frank Schmidt, University of Iowa

To better understand the relationships between admission test scores, high school GPA, parental income, first-year GPA (FYGPA) and second-year retention, we tested a meta-analytic path model. ACT scores, HSGPA, and SES had direct effects on FYGPA, but FYGPA fully mediated their effects on second-year retention.

Electronic Board #17

Identifying compromised items by subgroup item difficulties based on response times

Shu-chuan Kao, Pearson

The time sensitivity index is proposed to flag compromised items for Rasch-calibrated, computer-based tests by evaluating the parameter invariance assumption when item latency is considered. Compromised items are indicated

by short response times and correct responses. The feasibility of the proposed method will be demonstrated by simulated and imperial data.

Electronic Board #18

Evaluation of three types of DIF in multilevel mixture IRT models

Jungkyu Park, McGill University; Kwanghee Jung, Texas Tech University; Jaehoon Lee, Texas Tech University

This simulation study compares four different testing procedures using multilevel mixture item response model (MMIRT) in order to explore the best practice of DIF analysis when three different types of DIF — level-1 observed DIF, level-1 latent DIF, and level-2 latent DIF — exist in a test.

Electronic Board #19

Assessing M2 and RMSEA2 of Multidimensional Item Response Theory Models

Caihong Li, University of Kentucky; Hao Zhou, University of Kentucky; Michael Toland, University of Kentucky

This simulation study aimed to investigate the performance of M_2 and $RMSEA_2$ for polytomous data under multidimensional models. M_2 was found to have normal Type I error rates but unstable power. We hope to enlighten applied researchers on the usage of M_2 and $RMSEA_2$ when data is multidimensional polytomous.

Electronic Board #20

Comparing Groups of Correlation Matrices Using Fisher's z and Multiple Comparison

Jay Verkuilen, CUNY Graduate Center; Sydne McCluskey, CUNY Graduate Center

We consider visualization of groups of correlation matrices—such as encountered in group comparison or invariance studies—based on Fisher's z. Because the number of correlations becomes very large, we use multiple comparisons to provide an approximate probability calibration. A data example is provided.

Saturday, April 14, 4:05 – 6:05, 2018
Ambassador 3, Invited Session, E1

Measurement Problems – A look back to help us look ahead
Measurement Problems Session 1

Session Moderator: Mark Wilson, University of California, Berkeley (Immediate Past President)
Panelist: Michael Kolen, University of Iowa, NCME President 1999-2000
Panelist: Suzanne Lane, University of Pittsburgh, NCME President 2003-2004
Panelist: Laurie Wise, HumRRO, NCME President 2014-2015

History teaches the continuity of science; the developments of tomorrow have their genesis in the problems of today. Thus any attempt to look forward is well begun with an examination of unsettled questions. Since a clearer idea of where we are going smoothes the path into the unknown future, a periodic review of such questions is prudent. The present day, lying near the juncture of the centuries, is well suited for such a review. This article reports 16 unsolved problems in educational measurement and points toward what seem to be promising avenues of solution.

So begins Howard Wainer's 1993 article, Measurement Problems, a call "to begin to formulate the problems of our field" as we approached the end of the twentieth century. Twenty-five years later, another period review is prudent. In the spirit of the original article, we will devote three conference sessions to reflect on the most important problems in our field.

In this first session, a panel of NCME past presidents will consider the status of Wainer's original list of 16 unsolved problems; which have been solved, which remain, and what new challenges have emerged.

Saturday, April 14, 2018
4:05-6:05pm, Broadway I, Coordinated Sessions, E2

Measuring Essay Writing Competency in Europe using Human and Automated Scoring

Session Chair: André Rupp, Educational Testing Service
Session Chair: Stefan Keller, Fachhochschule Nordwestschweiz
Session Discussant: Mark Shermis, University of Houston-Clear Lake
Session Discussant: Olaf Koeller, University of Kiel

In this symposium we critically discuss the research design, methodology, and key findings from a large-scale longitudinal study for measuring EFL writing proficiency using digitally-delivered essays across six cantons in Switzerland and one federal state in Germany. Two different types of essays were used that required different types of argumentation and information synthesis from sources. The same learners provided responses at two different time points, set about one school term apart, and responded to instruments that measured related psychological competencies as well as questionnaires that captured class- and school-level context characteristics. The essay responses were then analyzed through a complex rating design produce high-quality human ratings and were then used to train and evaluate automated scoring models using state-of-the-art computational tools. The performance of these models was compared to existing models and scores were aligned to the CEFR as well as used in multi-level regression model to predict changes in writing competency across the two time points as a function of individual-, classroom-, and school-level factors. Findings demonstrate high human rating performance, interactions between prompt, population, and modeling approach, and complex relationships between factors at different levels. Practical recommendations for best research and development practices for like projects are provided.

Study Context and Overview
Stefan Keller, Fachhochschule Nordwestschweiz

Human Scoring
Jodi Casabianca-Marshall, Educational Testing Service

Automated Scoring
André Rupp, Educational Testing Service

Score Reporting
Johanna Fleckenstein, University of Kiel

Explanatory Modeling
Maleika Krueger, Fachhochschule Nordwestschweiz; Jennifer Meyer, University of Kiel

Saturday, April 14, 2018
4:05-6:05pm, Broadway II, Coordinated Sessions, E3

Considerations for Best Practices in Scale Development

Session Chair: Joseph Martineau, Center for Assessment
Moderator: Leslie Keng, Center for Assessment
Panelist: Andrew Middlestead, Michigan Department of Education
Panelist: Derek Briggs, University of Colorado Boulder
Panelist: Walter (Denny) Way, The College Board

Scaling is the means of translating each examinee's body of responses on a test into reported scores, and is often considered a relatively simple and routine task. However, in effective scaling, practitioners must resolve general indeterminacy from calibration, range indeterminacy from form-to-form differences in difficulty, and minimize problematic interpretations associated with imprecision, all while attaching desirable meaning to specific score points. Finding an optimal solution amidst technical and policy constraints is challenging, particularly in the absence of consensus guidance. Much of the existing guidance from one stakeholder perspective conflicts with guidance from another. In addition, research on this issue has followed a similar pattern in that it is conducted through a single stakeholder lens rather than a multifaceted perspective, often focusing on a solution to one challenge without addressing how the solution may limit the scale from the perspective of another challenge. Finally, there has been little research on the downstream effects of scaling decisions on score reporting and scale maintenance. This coordinated session uses an innovative format to initiate the process of developing a set of best practices for scale development through synthesis of policy, technical, and utility considerations combined with understanding of downstream effects.

Scale Development Guidance and Best Practices
Jennifer Dunn, Questar Assessment, Inc.

Policy Considerations in Scale Development
Jeffrey Hauger, New Jersey Department of Education

Technical Considerations in Scale Development
Gautam Puhan, Educational Testing Service; Neil Dorans, Educational Testing Service

Downstream Effects of Scaling Decisions on Stability and Fairness
Joseph Martineau, Center for Assessment

Saturday, April 14, 2018
4:05-6:05pm, Broadway III, Coordinated Sessions, E4

Towards Understanding the Facilitators and Inhibitors in Writing Tasks Containing Multimedia-Enhanced Stimuli

Session Chair: Young Kim, American Institues for Research
Session Chair: Peggy Carr, National Center for Education Statistics
Session Discussant: Yvonne Fuentes, University of West Georgia
Session Discussant: Jodi Davenport, WestEd

The launch of NAEP's transition to digital-based assessments across all subject areas raises a question of the role of multi-media features in assessments, including writing assessment, especially for fourth-graders. The purpose of the study was to investigate empirically whether features of writing tasks, including multimedia, can be systematically manipulated to make the writing tasks more accessible to students, especially to low-performing fourth-grade students, while remaining aligned with the NAEP writing framework. The study consists of two parts: (1) cognitive interviews to collect data on students' perceptions regarding facilitators and inhibitors present in the original and modified versions of two multi-media writing tasks and (2) a small-scale group administration of the writing tasks to assess students' actual performance on the original and modified tasks. The results of the study will help improve interpretation of data, augment the validity and utility of the assessment, and inform the development of multimedia-enhanced tasks. This symposium consists of five presentations reporting various aspects of the study and remarks by two discussants with expertise in multimedia features and plain language. The symposium will conclude with questions from the audience and answers from the presenters and discussants.

Role of multimedia features in NAEP writing assessment
Sheida White, National Center for Education Statistics

Design of cognitive interviews and group administration
Steven Hummel, American Institutes for Research

Analysis of cognitive interview and writing score data
Fran Stancavage, American Institutes for Research

Findings from cognitive interviews and group administration
Young Yee Kim, American Institutes for Research

Implications of study findings for the use of multimedia-enhanced stimuli in assessment
Jing Chen, National Center for Education Statistics

Saturday, April 14, 2018
4:05-6:05pm, Gershwin II, Individual Presentations, E5

Detecting Bad Things: Research on Cheating

Session Discussant: James Wollack, UW-Madison

Two Modifications of the Erasure Detection Index for Groups
Sandip Sinharay, Educational Testing Service

We suggest two new statistics for detecting potentially fraudulent erasures at an aggregate level. The statistics are modifications of the erasure detection index for groups (Wollack & Eckerly, 2017). The statistics are shown to have satisfactory Type I error rate and power for simulated data.

Detecting Groups of Examinees Involved in Test Collusion
Dmitry Belov, Law School Admission Council; James Wollack, University of Wisconsin-Madison

Test collusion (TC) is a large-scale sharing of test materials or answers to test questions. Because of potentially large groups involved, TC poses a serious threat to the validity of score interpretations. Proposed approach applies graph theory methodology to response similarity analyses for identifying groups while minimizing Type I error.

A new statistic for detecting aberrant response time patterns in large-scale assessments
Zhen Li, eMetric; Nathan Wall, eMetric; Huixing Tang, eMetric

Description of an easy-to-compute statistic for detecting examinee's aberrant response times in large-scale assessments. A simulation study and an empirical study were conducted to evaluate its performance. Results show that the new statistic performed equivalently well to van der Linden & Guo's (2008) Bayesian procedure, and reduced computation burden monumentally.

Use of Data Mining Methods to Detect Test Fraud
Kaiwen Man, University of Maryland College Park; Sandip Sinharay, educational testing service; Qian Yao, Educational Testing Service; Jeffrey Harring, University of Maryland College Park; Hong Jiao, University of Maryland College Park

Data mining methods have drawn considerable attention in diverse scientific fields. However, few applications have focused on test security research. In this study, various data mining methods for cheating detection have been explored with a common dataset from the Handbook of Quantitative Methods for Detecting Cheating on Tests.

Enhancing the Sensitivity of the J2 Model for Detecting Test Cheating
Gregory Hurtz, PSI Services LLC; John Weiner, PSI Services LLC

Test security is a major concern in high-stakes testing, and data forensics strategies are becoming increasingly utilized. We evaluate variants on the J_2 model of response similarity, demonstrating that using two variants together is more effective for detecting multiple patterns than more complex indices requiring probabilities from item response models.

A new method to detect aberrant erasures
Yuyu Fan, Fordham University; Joseph Grochowalski, College Board; Amy Hendrickson, College Board

We propose a method of erasure fraud analysis based on classical test theory and investigate its statistical power and type I error rate in a simulation study. Unlike previous simulation studies to detect erasure fraud, we simulate data to have a correlation between erasure corrections and examinee ability.

Saturday, April 14, 2018
4:05-6:05pm, Belasco, Individual Presentations, E6

Application and Evaluation of DCM

Session Discussant: Benjamin Shear, Colorado

Using Cognitive Diagnosis Modeling to Identify Students for Targeted Remediation
Xin Liu, Ascend Learning; Jennifer Brussow, Ascend Learning; Haiqin Chen, American Dental Association; Christine Mills, Ascend Learning

This study validates the utility of Cognitive Diagnosis Modeling in identifying students for targeted remediation by examining the accuracy of predicting a known outcome on a licensure exam. Furthermore, this study describes a practical implementation of the CDM-estimated indices in real testing practice such as setting CDM based benchmarks.

A Q-Matrix Validation Method for Continuous Response CDMs
Nathan Minchen, Rutgers, The State University of New Jersey; Jimmy de la Torre, The University of Hong Kong

Q-matrix validation methods exist for binary response but not continuous response cognitive diagnosis models. Recently proposed continuous response models require a new method. A flexible method is developed for a generalized continuous response model. Results from a simulation study and a real data example demonstrated the method's viability.

A Nationwide Cognitively Diagnostic Assessment Application in Natural Numbers
Lokman Akbay, Mehmet Akif Ersoy University / Turkey; Türker Toker, Uşak University / Turkey; Mehmet Kaplan, Artvin Çoruh University / Turkey; İbrahim Yıldırım, Harran University/ Turkey; Şerife Seviş, Middle East Technical Universiy / Turkey; Burcu Parlak, MoNE / Turkey; Erdinç Çakıroğlu, Middle East Technical Universiy / Turkey

This study aims to identify and develop the required attributes in the domain of natural numbers for elementary education. In the identification and validation process; psychometricians, curriculum experts, academicians in mathematics education, and teachers have participated in two workshops in which they identified the required attributes about natural numbers.

Investigating impact of Q-matrices on CDM of reading: Does curriculum alignment matter?
Clarissa Lau, University of Toronto; Megan Vincett, University of Toronto; Eunice Jang, University of Toronto

In order for feedback to effectively support students, it needs to be provided at an appropriate level of granularity. This study generated diagnostic reading profiles from provincial literacy assessment and compared the profiles with curriculum standards. Implications to current reporting policies are examined and discussed.

Determining an operationally appropriate level of grain size in Cognitive Diagnostic Models
Elizabeth Patton, University of North Carolina Greensboro; Alexandra Lay-Martin, University of North Carolina Greensboro; Robert Henson, University of North Carolina Greensboro

This research seeks to identify the impact attribute grain size has on item parameter estimation, model fit, and attribute profile pattern estimation. A variety of factors were investigated including number of attributes to be combined, method for redefining the Q-matrix, degree of correlation between attributes, and model chosen.

A Comparison of C-RUM and MIRT in Item Parameters and Classification Accuracy

Yanan Feng, Indiana University Bloomington; Dubravka Svetina, Indiana University Bloomington

This study investigates how comparable are multidimensional item response theory (MIRT) and compensatory reparameterized unified models (C-RUM) in terms of item parameters and classification accuracy. Specifically, we are interested in how accurate the classification will be if we retrofit cognitive diagnostic models (CDM) to IRT-constructed assessments, or vice versa.

Saturday, April 14, 2018
4:05-6:05pm, Plymouth, Individual Presentations, E7

Investigating Fit

Session Discussant: Scott Monroe, University of Massachusetts Amherst

Investigating the Practical Impact of Model Misfit in IRT

Hwanggyu Lim, University of Massachusetts Amherst; Minjeong Shin, American Institutes for Research; Ah-Young Shin, American Institutes for Research

This study suggests a practical method for evaluating the impact of model data misfit by estimating the variance of equating via a simulation study as well as an application to the real data from a large-scale educational assessment.

Methods for Improving the Goodness-of-fit By Considering Responses and Response Time

Heru Widiatmo, ACT, Inc.; Lisa Gawlick, ACT, Inc.

The Effective Response Time method, which uses both responses and response times, is compared to and combined with Person-Fit Statistics that use only responses to find an optimal method for improving the goodness-of-fit. The 2- and 3-PL IRT models are used to calibrate items and to evaluate the results.

Evaluating a modified nonparametric procedure to assess parametric IRT model fit

Adrienne Sgammato, Educational Testing Service; John Donoghue, Educational Testing Service

The quality of parametric IRT estimates within a nonparametric framework is evaluated in this simulation study using an adaptation of Douglas and Cohen's (2001) using a different source of item parameters. Results suggest that Type I error is fairly well controlled and power is >= .5 for larger sample sizes.

The Effects of Collapsing Ordered Categorical Variables on Tests of Measurement Invariance

Yuan-Ling Liaw, University of Oslo Centre for Educational Measurement; Leslie Rutkowski, University of Oslo Centre for Educational Measurement; Dubravka Svetina, Indiana University-Bloomington

In measurement models, ordered categorical outcomes are collapsed for substantive reasons or because of sparse cells. Using empirical and simulated data, we examine the impact of collapsing categories in a multi-group context. The impact on model fit, parameter estimates, standard errors, and scale reliability under different collapsing decisions are reported.

A stepwise procedure for determining measurement invariance using IRT item fit

Janine Buchholz, German Institute for International Educational Research (DIPF); Johannes Hartig, German Institute for International Educational Research (DIPF)

The Programme for International Student Assessment (PISA) recently introduced an IRT-based item-fit approach to testing measurement invariance suitable for large numbers of groups. To overcome its limited power for detecting cross-group variation in item discrimination, this study reports promising evidence resulting from a stepwise relaxation of item constraints across groups.

Generalized S-X^2 under different response categories with non-normal latent trait using IRT.

Sunil Lamsal, Pearson VUE; Joe Betts, Pearson VUE

The performance of item-fit indices for polytomous item response models is important for gauging the utility of items in operational testing programs. This study focuses on generalized S-procedure to assess the item fit statistics under different response categories and non-normal latent trait using partial credit and generalized partial credit models.

Saturday, April 14, 2018
4:05-6:05pm, Manhattan, Individual Presentations, E8

New Research on Multidimensional IRT

Session Discussant: Richard Schwarz, Educational Testing Service

Applications of Multivariate Techniques to Measure Content Structure with Multidimensional IRT
Quinn Lathrop, Pearson Advanced Computing and Data Science Lab

This work demonstrates the use of multivariate techniques, such as Factor Analysis, to make inferences about the structure of content in online learning systems. By leveraging the covariance matrix output from multidimensional IRT, these tools can provide recommendations to improve the learning experience of students.

Examining Compensation at the Item-level in a Multidimensional Assessment
Xinchu Zhao, University of South Carolina; Brian Habing, University of South Carolina

The purpose of this study is to compare the compensatory, noncompensatory and a new MIRT model, named the Rotatable Asymmetric Variable Compensation Model on a real data set. The log-likelihood given by the fitted models are compared at item level on the real data, and data simulated to be similar.

Underfitting 2-Dimensional Data with the Generalized Graded Unfolding Model: Item Structure Effects
James Roberts, Georgia Institute of Technology; Jordan Sparks, Georgia Institute of Technology; David King, Pacific Metrics

This research investigates the effects of multidimensional item structure (simple versus complex) on the characteristics of estimates that result from misapplying the (unidimensional) generalized graded unfolding model to 2-dimensional data. The role of alternative parameter estimation methods on the resulting direction of best measurement is also studied.

Application of Multidimensional IRT to a Test of K-12 English Language Proficiency
Li Cai, University of California, Los Angeles; Mark Hansen, University of California, Los Angeles

We describe the application of multidimensional item response theory models in the calibration and scoring of tests of English Language Proficiency (ELP) recently developed by and now used in the ELPA21 consortium.

Saturday, April 14, 2018
4:05-6:05pm, Ambassador II, Coordinated Sessions, E9

Fairness in Testing ELs and ELs with Disabilities: Research, Implementation, and Policy

Session Chair: Edynn Sato, Sato Education Consulting, LLC
Session Discussant: Martha Thurlow, National Center on Educational Outcomes

Assessment results, particularly high-stakes results, can promote or limit students' subsequent opportunities and the degree to which they are able to thrive in and contribute to society. Students who are English learners (ELs) and English learner students with disabilities (ELSWDs) in the U.S. face significant testing challenges in that they are not yet proficient in the language of assessment (i.e., English), have cultural orientations that may impact their meaning-making and related performance, and/or have disabilities that affect their engagement with assessment tasks and their capacity to demonstrate fully what they know and can do. The papers in this session reflect the complexities of fairly and validity assessing ELs and ELSWDs, and they are intended to inform test developers, psychometricians, policymakers, and educators concerned with issues relevant to the fair and valid testing of these diverse learners. From the following perspectives: (a) socio-cultural and accessibility; (b) development and implementation; (c) psychometric; and (d) policy and legal, presenters will discuss relevant research and practice and offer heuristics intended to augment our understanding and practices related to fairness in testing our ELs and ELSWDs so that assessment outcomes are accurate and meaningful and support students' opportunities and success in school and at work.

Psychometric Perspectives on Fairness in English Language Proficiency Assessments
Nami Shin, CRESST

Policy Considerations: Assessing Els and ELSWDs for Classification and Accountability
Margaret Ho, CRESST

Fairness in Testing: Assessment Development and Implementation
Michelle McCoy, CRESST

Sunday, April 15, 2018
10:35 – 12:05, Majestic 2, Invited Session, F1

The Positive Impact of Assessment

Session Moderator: Brian Gong, National Center for the Improvement of Educational Assessment
Panelist: Joanna Gorin, Educational Testing Service
Panelist: Margaret Heritage, WestEd
Panelist: James Pellegrino, University of Illinois at Chicago

The theme of the 2018 NCME Conference is *Here and There and Back Again: Making Assessment a Stronger Force for Positive Impact on Teaching and Learning*. That conference theme is based on two of the four "directions" for NCME that Randy Bennett laid out in his initial president's message:

- Encourage research and development that makes assessment a stronger force for positive impact on teaching and learning;
- Encourage and promote the positive influences of classroom assessment on measurement, and the positive influences of measurement on classroom assessment

In this session, panelists will draw on their distinguished backgrounds and areas of expertise to address both ways that assessment has been a positive impact on teaching and learning and ways that it could become a more positive influence in the future. Panelists will consider assessment in broad terms, addressing various forms of assessment processes and practices intended to generate information for a variety of purposes and uses.

Through their discussion, panelists will also address barriers that have prevented assessment from become a stronger force for positive impact in the past and reflect on how those barriers can be overcome in the future.

Sunday, April 15, 2018
10:35am - 12:05pm, Ambassador III, Individual Presentations, F2

Technology-Based Asssesment: Tests, Items, and Methods

Session Discussant: Kirk Becker, Pearson

Modeling Slipping Effect in a Large Scale Assessment with Innovative Item Formats

Ismail Cukadar, Florida State University; Salih Binici, Florida Department of Education

This study employs the 4PL-IRT model to account for unexpected incorrect responses or slipping effect in a large scale Algebra 1 assessment. It investigates whether modeling the misfit at the upper asymptote has any practical impact on student ability estimates. A simulation study was also conducted to support the findings.

Developing Authentic Digital Math Assessments

Laurie Davis, ACT, Inc.; Kristin Morrison, ACT, Inc.; Yile Zhou, ACT, Inc.

This study evaluates the feasibility of a digital assessment item format for mathematics with high construct fidelity that allows a student to solve and show their work for an item on a tablet using a digital pen and compares it to paper- and type-written response formats.

Instructional validity of a new video-based assessment for measuring teachers´ instructional performance

Christiane Kuhn, Johannes Gutenberg-University Mainz (Germany), Department of Business and Economics Education; Olga Zlatkin-Troitschanskaia, Johannes Gutenberg-University Mainz (Germany), Department of Business and Economics Education; Sebastian Brückner, Johannes Gutenberg-University Mainz (Germany), Department of Business and Economics Education; Hannes Saas, Johannes Gutenberg-University Mainz (Germany), Department of Business and Economics Education

We evaluated the instructional validity of a video-based assessment for measuring teachers´ performance using 48 expert ratings and 42 cognitive interviews with teachers. We found evidence of instructional validity, as the ratings, the interviews, and their relationships correspond in showing the relevance and suitability of the assessment for instructional practice.

Validity Inferences for Different Types of Technology-Enhanced Items

Angela Hochstetter, Minnesota Department of Education; Ann Page, Minnesota Department of Education; Yu-feng Chang, Minnesota Department of Education; Kevin Cappaert, Pearson

We propose that HotSpot (HS) and Match-Table-Grid (MTG) items more accurately measure student ability than multiple-response items. We hope to provide guidance for assessment practitioners to effectively select the most appropriate technology-enhanced item types based on the concept assessed.

Optimizing Partial Credit Scoring for Multi-Component Technology Enhanced Items

Shuqin Tao, Curriculum Associates

This study proposes an optimal partial credit scoring approach and applies it to a variety of multi-component technology-enhanced item types. Findings will shed light on its effectiveness, validity and applicability and provide evidence on its improvement upon the one-size-fits-all scoring approach as currently implemented in PARCC.

Designing a More Authentic Science Assessment Environment: A Virtual Science Laboratory

Timothy Fiser, Educational Testing Services; Shu-Kang Chen, Educational Testing Services; Raymond De Hont, Educational Testing Services; Katherine Castellano, Educational Testing Services; Lei Liu, Educational Testing Services; Delano Hebert, Educational Testing Services; Kenneth Llort, Educational Testing Services

The Virtual Science Laboratory (VSL) prototype was developed to assess what students know and can do in science. The VSL is an open-ended, virtual 3D laboratory with touch interactive supplies and scientifically accurate simulated phenomena. Cognitive labs indicated that students can successfully design and conduct virtual investigations and communicate conclusions.

Sunday, April 15, 2018
10:35am - 12:05pm, Melville, Individual Presentations, F3

New Directions for Multilevel Models

Session Discussant: Dena Pastor, James Madison University

The Multilevel Measurement Model for Partially Clustered Individuals

Luping Niu, The University of Texas at Austin; Tasha Beretvas, The University of Texas at Austin

The present study introduced the multilevel measurement model (MMM) for partially clustered individuals, and assessed how estimation of the proposed MMM performed under different conditions. The results could inform researchers of designing and modeling partially clustered data with IRT-based test items using appropriate settings

Score-based Tests for Comparing Treatment Effects in Multilevel Models

Ting Wang, The American Board of Anesthesiology; Edgar Merkle, University of Missouri; Joaquin Anguera, University of California, San Francisco; Brandon Turner, The Ohio State University

It is often difficult to compare the magnitude of treatment effects in clustered data. This is because unequal variance components across treatments can easily be mistaken as differences in the effects of interest. We utilize a recently-proposed family of score-based tests to distinguish between these two issues.

Determining Predictor Relative Importance in Explanatory Multilevel IRT Models

Luciana Cancado, University of Wisconsin-Milwaukee; Razia Azen, University of Wisconsin-Milwaukee

Explanatory multilevel IRT models allow the inclusion of predictors at various levels when estimating latent traits. Once model predictors are selected, one might want to rank their relative contributions. This simulation study evaluates the use of Dominance Analysis for determining predictor relative importance in Kamata's (2001) three-level IRT model.

Computing Test Score Distributions with the Hierarchical Rater Model

YoungKoung Kim, The College Board; Tim Moses, The College Board; Lawrence DeCarlo, Teachers College Columbia University

An approach to computing test score distributions is presented for constructed response (CR) items scored by raters. The estimation of test score distributions using the Hierarchical Rater Model (HRM) with the Lord-Wingersky algorithm is described and results are compared to those from simpler IRT models without rater effects.

FIML estimation of LATE through latent RD analysis with an MH-RM algorithm

Monica Morell, University of Maryland; Ji Seung Yang, University of Maryland

A Metropolis-Hastings Robbins-Monro (MH-RM) algorithm is implemented in R and evaluated via Monte Carlo simulations to obtain an unbiased full-information maximum likelihood (FIML) local average treatment effect (LATE) in the regression discontinuity (RD) design where an item response theory model (IRT) is used for the latent treatment assignment variable.

Sunday, April 15, 2018
10:35am-12:05pm, Majestic I, Coordinated Sessions, F4

Students' Use of Response Time, Testing Behavior, and Performance in Digitally-Based Assessments

Session Chair: Young Yee Kim, American Institutes for Research
Session Chair: Markus Broer, American Institutes for Research
Session Discussant: Ryan Baker, University of Pennsylvania

In recent years, more assessments including the National Assessment of Educational Progress (NAEP) are transitioning from paper-based assessments (PBAs) toward digitally-based assessments (DBAs). The transition to DBA permits the collection of detailed timing data on students' test taking behaviors. Automatically collected behavioral data provide a rich data source to examine the relationship between students' testing behavior and performance from various aspects. This symposium features three separate studies investigating the relationship between students' testing behavior and performance, using the 2016 NAEP mathematics grade 8 pilot DBA administered to a nationally representative sample of about 12,000 students. The first study examines the relationship between students' time management strategies and performance.

The second study focuses on the issue of the effectiveness of extended time accommodation (ETA), by analyzing the relationship between ETA and performance of students with ETA.

The third study deals with a huge issue in large-scale, low stakes test, such as NAEP, rapid-guessing. This study uses growth mixture models (GMM) to identify rapid-guessers.

Exploring the relations between students' time management strategies and test performance
Fusun Sahin, American Institues for Research; Qin Lu, University of Kansas; Tiago Calico, American Institues for Research

The Extended Time Accommodation (ETA) and Performance of Students with ETA
Young Yee Kim, American Institues for Research; Ruhan Circi, American Institues for Research

Identifying Rapid-Guesser Using Growth Mixture Models
Xiaying Zheng, American Institutes for Research; Tanesia Beverly, University of Conneticut; Young Yee Kim, American Institues for Research

Sunday, April 15, 2018
10:35am-12:05pm, Gershwin II, Coordinated Sessions, F5

Using an Assessment Use Argument in developing, using, and justifying K-12 assessments

Session Chair: Lyle Bachman, University of California, Los Angeles
Session Chair: H. Gary Cook, University of Wisconsin-Madison

Assessments play an important role in K-12 education: providing measures that inform decisions about students, teachers, and programs. These decisions range from very high-stakes summative decisions to relatively lower-stakes formative decisions that are aimed at improving instruction and learning. Given the importance of these decisions, practitioners—test developers and test users—need to be able to evaluate and demonstrate the relevance, utility, and consequences of their assessments.

Bachman and Palmer (2010) have developed an innovative approach to evaluating the quality of an assessment that they call "assessment justification". This shifts the focus from the validation process itself to the purpose for which this process is intended—justifying the uses of assessments to stakeholders. Their approach specifies explicit links from assessment performance to interpretations, decisions, and consequences, and thus extends current argument-based approaches to validation beyond interpretations to assessment use.

The presentations in this coordinated session provide an overview of Bachman and Palmer's approach, along with four examples of how this approach is being applied in both large-scale and classroom assessments in K-12 to provide stronger positive impact on learning and teaching. The session will conclude with an open discussion between the presenters and the audience

Justifying the uses of assessments
Lyle Bachman, University of California, Los Angeles

Facilitating Assessment Use as the Guiding Principle in a Large-Scale Assessment Program
Dorry Kenyon, Center for Applied Linguistics

Test-based decisions that inform teaching and learning of K-12 English Learners
Ahyoung Alicia Kim, University of Wisconsin-Madison

Justifying the use of integrated assessments of language and content
Lorena Llosa, New York University

Using an assessment use argument to guide classroom-based assessments
Barbara Damböck, Akademie Dillingen, Germany

Sunday, April 15, 2018
10:35am - 12:05pm, Belasco, Individual Presentations, F6

Issues in Linking and Equating

Session Discussant: Sonya Powers, ACT

Standard errors of IRT true-score equating: a multiple-imputation approach

Zhonghua Zhang, Melbourne Graduate School of Education, University of Melbourne

This study evaluated a multiple imputation-based procedure for estimating the standard errors of IRT true-score equating coefficients. The simulation results indicated that this multiple-imputation based method could be a practically viable alternative to the bootstrap method and the delta method when the calibration sample size was reasonably large.

A simple parametric procedure for detecting drift in anchor items

Xi Wang, Measured Progress; Louis Roussos, Measured Progress

Building on a previously proposed estimator for detecting drifting items, two methods are proposed for augmenting it with a standard error estimator. The enhanced statistical procedure is evaluated in a simulation study.

Challenges in IRT-Linking with Longitudinal Designs

Luise Fischer, University of Bamberg; Timo Gnambs, LIfBi; Theresa Rohm, University of Bamberg; Claus Carstensen, University of Bamberg

Findings of the present simulation study suggest negligible differences in mean bias and mean error among separate calibration IRT-linking methods with regard to number of anchor items, sample size, model fit, and proficiency variance. In contrast, concurrent calibration seems less robust with regard to challenges inherent to longitudinal link designs.

The Effect of Anchor Construction on Test Score Equating

Yongmei Zhang, Beijing Academy of educational sciences; Jiaqi Wang, Beijing Academy of Educational Sciences; Meijuan Li, Beijing Academy of Educational Sciences; Yi Tian, Beijing Academy of Educational Sciences; Yi Hao, Beijing Academy of Educational Sciences; Hongqi Chu, Beijing Academy of Educational Sciences

A simulation study was conducted to investigate the effect of content and statistical representativeness of an anchor on test score equating in K-12 related examinations. The results show that miditests perform as well as or better than minitests and appear to be relatively robust to large differences in group ability.

Impact of Degrees of Postsmoothing on Long-Term Equated Scale Score Accuracy

Stella Kim, The University of Iowa; YoungKoung Kim, The College Board; Tim Moses, College Board; Caiyan Zhang, The College Board; Judit Antal, The College Board

The proposed study attempts to examine the long-term implications of the degrees of postsmoothing in equating for equated scale score accuracy. The study examined several factors based on simulation, including the number of equating chains, degrees of postsmoothing and test length.

Impact of score distributions on precision of chained equipercentile equating

Yanlin Jiang, Educational Testing Service

The study explores potential changes in equating precision with chained equipercentile when score distributions vary under the common-item design. IRT 2PL simulated data will be used and the results of equating precision under various score distributions are provided and evaluated in this study.

Sunday, April 15, 2018
10:35am-12:05pm, Plymouth, Coordinated Sessions, F7

Exploring Properties, Issues, and Solutions with Estimating Student- and Aggregate-Level Growth Measures

Session Chair: Katherine Castellano, Educational Testing Service
Session Discussant: Scott Monroe, University of Massachusetts Amherst

Estimating student-level growth to describe student progress and aggregating student growth measures for educator evaluations or school/district accountability remains a hot topic for state departments of education as they weigh competing statistical, political, and logistical priorities. This session highlights a series of papers that investigate properties (e.g., estimation accuracy and reliability) and issues (e.g., mode-effects) at both the student- and aggregate-levels. The papers all consider the popular Student Growth Percentile (SGP) measure or the mean/median of this statistic at the educator, school, or district level to some extent. However, they also consider a range of other growth measures, including residual gains (both observed and Expected-a-Posteriori estimates), Student Learning Objectives (SLOs), and value-added measures.

Through these rigorous, statistical studies the authors not only make substantial contributions to the body of literature on student growth but also underscore the importance of practitioners' concerns, including communicability and utility.

SGP Measurement Error Correction: An Empirical Investigation of the Ranked SIMEX Approach
Damian Betebenner, Center for Assessment; Adam van Iwaarden, Center for Assessment

When does Conditioning on Multiple scores Improve the Accuracy of Residual Gains?
Katherine Castellano, Educational Testing Service; Daniel McCaffrey, Educational Testing Service

Mode of Assessment Administration and its Impact on SGP Calculations
Kathleen Flanagan, Massachusetts Department of Elementary-and-Secondary-Education; Damian Betebenner, Center for Assessment

The Intertemporal Variability of Student Learning Objective Ratings as Measures of Growth
Derek Briggs, University of Colorado; Rajendra Chattergoon, University of Colorado; Amy Burkhardt, University of Colorado

Should Aggregate Student Growth Measures Be Used to Measure Educator Performance?
Daniel McCaffrey, Educational Testing Service; J.R. Lockwood, Educational Testing Service; Katherine Castellano, Educational Testing Service

Sunday, April 15, 2018
10:35am-12:05pm, Manhattan, Coordinated Sessions, F8

Item Difficulty Modeling: Lessons Learned and Future Directions

Session Chair: Jeffrey Steedle, ACT
Session Discussant: Kristin Morrison, ACT

Item difficulty modeling (IDM) involves the application of statistical prediction models to examine the relationship between item features and item difficulty, often with the goal of identifying which item features are the most important predictors. With that knowledge, newly developed items can better target certain levels of difficulty (e.g., for automatic item generation), and sources of construct-irrelevant variance can be identified and minimized. Moreover, results can support validity arguments by providing evidence that examinees apply the intended knowledge and skills as defined in achievement level descriptors. This coordinated session includes papers that provide background on IDM and compare methods, demonstrate IDM for improving item quality in operational assessment programs, use IDM to improve the transparency of inferences made about examinee ability, illustrate the consequences of erroneous item difficulty predictions on ability estimation, and examine the effect of performance feedback on estimates of item difficulty. In all, these papers broaden understanding of what makes items easy or difficult and advance methodology for gaining such knowledge.

Item Difficulty Modeling: Research Methods in Test Development
Susan Embretson, Georgia Institute of Technology

Using Item Difficulty Modeling to Improve Item and Test Quality: An Illustration
Steve Ferrara, Measured Progress; Jeffrey Steedle, ACT; Roger Frantz, Questar

Including Student Engagement Variables into Item Difficulty Models: An Exploratory Study
Kristen Huff, Curriculum Associates; Dan Mix, Curriculum Associates; Christine Zanchi, Curriculum Associates

Impact of Parameter Imprecision on Ability Estimation Under a CAT Delivery Model
Jonathan Weeks, Educational Testing Service; Isaac Bejar, Educational Testing Service

Item Difficulty Modeling of Fluid Reasoning on the Woodcock Johnson Test
Clifford Hauenstein, Georgia Institute of Technology; Susan Embretson, Georgia Institute of Technology

Sunday, April 15, 2018
10:35am-12:05pm, Ambassador II, Coordinated Sessions, F9

Boundary-pushing innovations in the assessment of English language learners, co-sponsored with AERA-IAEA

Session Chair: Joni Lakin, Auburn University
Session Discussant: Mikyung Wolf, Educational Testing Service

English language learner (ELL) students are a heterogeneous population of students that pose unique challenges to test development and administration. In response to these challenges, researchers have developed innovative and boundary-pushing solutions that expand our knowledge of assessment for ELL students as well as the general student population. The goal of this session is to highlight methods recently developed to enhance the validity and fairness of assessments for ELL students in the K-12 testing context. The goal of each line of research is to enhance the quality of information educators have about the instructional needs of their ELL students. Each assessment innovation also has implications for how we design all assessments with the principles of Universal Design. This session will highlight several innovations in assessment, particularly in the use of picture-based or nonverbal assessment approaches. **This session is co-sponsored by the *Inclusion and Accommodation in Educational Assessment* Special Interest Group of AERA.**

Designing and Evaluating Illustrations for a National Next Generation Mathematics Assessment
Magda Chia, Stanford University; Rachel Kachchaf, Smarter Balanced Assessment Consortium; Guillermo Solano-Flores, Stanford University

Universal Design, Fairness, and Pictorial Reasoning Assessments
Joni Lakin, Auburn University

ONPAR: A Multisemiotic Assessment Design for ELL Students
Moni McGlone, University of Wisconsin-Madison; Rebecca Kopriva, University of Wisconsin-Madison; Kyle Schultz, University of Mary Washington

Bilingual Content Assessments for ELL Students
Alexis Lopez, Educational Testing Service

Sunday, April 15, 2018
2:45 – 4:15, Majestic 2, Invited Session, G1

Measurement Problems – A look back to help us look ahead
Measurement Problems Session 2

Session Moderator: Henry Braun, Lynch School of Education, Boston College
Panelist: Karen Barton, Edmentum
Panelist: Li Cai, CRESST
Panelist: Jimmy de la Torre, The University of Hong Kong
Panelist: Chris Han, Graduate Management Admission Council (GMAC)
Panelist: Alina von Davier, ACTNext

I limit myself to educational measurement problems within a particular context--specifically, measurements that lead to a decision and a consequence. Thus, this discussion is not aimed at a process analogous to, let's say, measuring someone's height: A measurement is made, the result is reported, and that's it. The process I concern myself with here is more like that in the measurement of weight. There are standards associated with good health that are connected to weight and actions that can be taken to affect weight, and the success of those actions can be assessed. I believe that the lion's share of the educational situations requiring measurement is of the sort that suggests an action and has an outcome.

With this boundary stated, let us begin. (Wainer, 1993)

In this second Measurement Problems session devoted to formulating the problems of our field panelists will examine unsettled questions that vex us today and identify new measurement challenges that are emerging or are likely to emerge in a world in which the walls between assessment and instruction/learning are being broken down; a world of personalized instruction with a focus on the individuals' learning and growth.

Sunday, April 15, 2018
2:45-4:15pm, Ambassador III, Coordinated Sessions, G2

Tackling practical issues in small sample scaling and equating

Session Chair: Joshua Goodman, NCCPA
Session Discussant: Mark Raymond, National Board of Medical Examiners

Examination programs aim to maximize score fairness, security, and examinee friendliness (e.g., quick reporting, low-cost, frequent administrations). In large-scale testing programs, shrewd test design and application of robust psychometric methods are used to ensure each of these above listed factors is adequately addressed. However, there many highly specialized occupations, professions, or practices that require passing a test as a precursor to certification or licensure. Often the people working in these fields are limited in number, thus the volume of test-takers for any given administration are too small to safely apply large-scale psychometric methods. Low-volume programs would still like to maximize the score-fairness, security, and examinee-friendliness, but the limitation imposed by sample sizes means these certification/licensure organizations often have more limited options when addressing these factors. This session explores the challenges of scaling and testing in programs where sample sizes are small, focusing specifically which methods are most promising. These studies address both raw score equating methods developed especially for small samples as well as scaling using the Rasch model.

Effect of Sample Size on Common-Item Equating using the Dichotomous Rasch Model
Justin Gregg, CareSource; Michael Peabody, American Board of Family Medicine; Thomas O'Neill, American Board of Family Medicine

Investigating the classification accuracy of Rasch equating with very small samples
Andrew Dwyer, American Board of Pediatrics; Robert Furter, American Board of Pediatrics

Investigating Repeater Effects on Small-Sample Equating: Include or Exclude?
Hongyu Diao, University of Massachusetts Amherst; Lisa Keller, University of Massachusetts Amherst

Equating with small and unbalanced designs
Fen Fan, NCCPA; Joshua Goodman, NCCPA; Andrew Dallas, NCCPA

Sunday, April 15, 2018
2:45-4:15pm, Melville, Coordinated Sessions, G3

Using Repeater Data to Inspect Quality and Security in Continuous Mode Testing

Session Chair: Alvaro Arce, Pearson
Session Discussant: Jeffrey Steedle, ACT

This coordinated session brings four papers that address critical development and research in data quality and security in ongoing testing. Each paper uses retake examinees test scores to inspect quality and security of test administrations in contexts ranging from K-12 testing to College admission and Licensure and Certification Testing. Collectively, the papers provide methodological enhancements to current methods to scrutinize repeater data in data quality inspections. The presentations will include key takeaways for testing programs operating with data quality plans and recommendations to practitioners on ways to inspect data quality and security with repeater data.

Taking a CBT in Continuous Environment Twice: Dealing with Test Exposure Control
Avi Allalouf, National Institute for Testing Evaluation; Tony Gutentag, National Institute for Testing Evaluation; Marina Fronton, National Institute for Testing Evaluation

Using Retake Examinee Test Data to Monitor Data Quality in Continuous Testing
Alvaro Arce, Pearson; Suleyman Olgar, FLDOE Postsecondary Assessment; Lauren White, FLDOE Postsecondary Assessment; Leah Kaira, Evaluation Systems group of Pearson

The Modified Signed Likelihood Ratio Test and Its Application to Repeater Data
Sandip Sinharay, Educational Testing Services

Using Data Analytics to Flag Potential Retester Misconduct
Anna Topczewski, GED Testing Service

Sunday, April 15, 2018
2:45-4:15pm, Majestic I, Coordinated Sessions, G4

Assessments of Collaborative Problem Solving and Implications for PISA 2015

Session Chair: Qiwei He, Educational Testing Service
Session Chair: Maida Mustafić, University of Luxembourg
Session Discussant: Matthias von Davier, National Board of Medical Examiners

Collaborative problem solving (CPS) is a critical and necessary skill in educational settings and workforce. The assessment of CPS that was first introduced in the Programme for International Student Assessment (PISA) 2015 focuses on the cognitive and social skills related to problem solving in collaborative scenarios. This symposium addresses the assessment of CPS from complementing perspectives and thereby delineate the future relevance of computer-based assessment for collaboration, for the future of education, and for large-scale educational assessments. Taking the CPS assessment in PISA 2015 as an example, this symposium aims at embracing the mutual impact of new collaboration conceptions on the development of computer technologies to assess and understand collaboration and vice versa. The theme sheds light on various aspects, including the development of new methodologies in measuring CPS skills, the development of new assessment instruments, the analysis of collaboration data and the implications of research for international large-scale educational assessments. This session will highlight, through four connected papers augmented by a discussant, issues around computer-based assistance of collaborative behavior, the computer based preconditions for collaboration, methodological approaches as well as analyses of different views on collaborative behaviors.

Assessing Collaborative Problem Solving through Conversational Agents
Art Graesser, University of Memphis

Producing a Reliable Collaborative Problem Solving Scale in PISA
Qiwei He, Educational Testing Service

An Overview: Collaborative Problem Solving in Large-Scale Assessments
Samuel Greiff, University of Luxembourg; Maida Mustafić, University of Luxembourg

Intuitive Use of Technological-Support-Kit Fosters Problem-Solving Processes in Human-to-Human Collaboration
Inga Bause, Leibniz-Institut für Wissensmedien, Tübingen, Germany; Irina Brich, Leibniz-Institut für Wissensmedien, Tübingen, Germany; Ann-Katrin Wesslein, University of Tuebingen, Tübingen, Germany; Friedrich Hesse, Leibniz-Institut für Wissensmedien, Tübingen, Germany

Producing a Reliable Collaborative Problem Solving Scale in PISA
Hyo Jeong Shin, Educational Testing Service; Mary Louise Lennon, Educational Testing Service; Haiwen Chen, Educational Testing Service; Matthias von Davier, National Board of Medical Examiners

Validating PISA Collaborative Problem Solving by Face-to-Face, Self- and Teacher-Report Measures
Katharina Herborn, University of Luxembourg; Maida Mustafić, University of Luxembourg; Samuel Greiff, University of Luxembourg

Sunday, April 15, 2018
2:45-4:15pm, Gershwin II, Individual Presentations, G5

Reimagining Adaptive Testing

Session Discussant: Liru Zhang, DE DOE

A New Approach: Mixed Computerized Adaptive Multistage Testing
Anthony Raborn, University of Florida; Halil Ibrahim Sari, Kilis 7 Aralik University

Computerized adaptive testing and computerized multistage testing are two popular versions of adaptive testing with their own strengths and weaknesses. This study proposes and investigates a combination of the two procedures designed to capture these strengths while minimizing the weaknesses, provisionally named mixed computerized adaptive multistage testing.

Multidimensional Testlet Adaptive Testing under a Higher-order Structure Design
Jing-Ru Xu, Pearson; Joe Betts, Pearson

This research explored the application of a higher-order testlet-based model under a multidimensional adaptive testing (MAT) context. Designs under various conditions were simulated and investigated. The results showed the new design leads to an increase in estimation precision with computational convenience than a general testlet MAT given different evaluation criteria.

Fully Adaptive Multistage Testing: A Highly Efficient and Controlled Adaptive Testing Model
Xinrui Wang, Pearson VUE; Xiao Luo, National Council of State Boards of Nursing

Besides test efficiency, content quality is also critical for large-scale high-stakes exams. This study introduces a new testing model that provides test developers high efficiency and control over the test. The operational superiority of this model, comparing to computerized adaptive testing and multistage testing, is evidenced by a simulation study.

Detecting Misconceptions and Estimating Ability Simultaneously: A Hybrid Computerized Adaptive Testing Framework
Yawei Shen, The University of Georgia; Yu Bao, The University of Georgia; Shiyu Wang, The University of Georgia; Laine Bradshaw, The University of Georgia

This study develops a hybrid design framework for computerized adaptive testing based on the Scaling Individuals and Classifying Misconceptions model, which can efficiently detect students' misconceptions and estimate a latent ability simultaneously. The proposed designs are evaluated through simulation studies and show the corresponding power and advantages of the design.

A New Concept of Computerized Adaptive Testing: Global Adaptiveness in Administration Procedures
Jyun-Hong Chen, National Sun Yat-sen University; Hsiu-Yi Chao, National Chung Cheng University; Ching-Lin Shih, National Sun Yat-sen University

This study introduces the concept of global adaptiveness that pursues adaptive testing within and between examinees tests, rather than just within single item administration, to improve CAT's efficiency. Through simulation studies, the results indicated that CAT with global adaptiveness always yields more precise trait estimates than that with traditional CAT.

Sunday, April 15, 2018
2:45-4:15pm, Belasco, Individual Presentations, G6

Approaches to Decisions/Classification

Session Discussant: Nathan Dadey, NCEIA

Comparing Decision Errors in Measurement Decision Theory with Rasch Scoring
Andrew Jones, American Board of Surgery; Jason Kopp, American Board of Surgery

Measurement decision theory (MDT) is a model specifically designed to make classification decisions for examinees. Recently, MDT has been used in operational settings, necessitating more research comparing MDT to traditional measurement models. This research provided greater understanding about MDT probabilities and decision errors, and comparisons to traditional Rasch measurement.

Measuring Reliability of Student Mastery Classifications at Multiple Levels
William Thompson, University of Kansas - Dynamic Learning Maps; Amy Clark, University of Kansas - Dynamic Learning Maps; Brooke Nash, University of Kansas - Dynamic Learning Maps

Providing evidence of reliability is critical for operational assessments. For diagnostic assessments, reliability is typically presented at the attribute level. In K-12 settings, this is often insufficient, as results must be aggregated for accountability purposes. This study demonstrates how reliability can be estimated for aggregated attributes within a diagnostic assessment.

The Effects of Stakes on Psychometric Decisions
Joseph Grochowalski, The College Board; Yuyu Fan, Fordham Univeristy; Amy Hendrickson, The College Board

Test development decisions are often based on low-stakes administration data, but little is known about how test characteristics change under high-stakes administration. We examine the effects of stakes on item performance, examinee performance, and test construction by comparing results from pseudo equivalent low- and high-stakes samples.

Characterizing classification accuracy using posterior densities from multidimensional IRT scoring
Mark Hansen, University of California, Los Angeles; Li Cai, University of California, Los Angeles

We examine an approach for quantifying the probability of correct classification at both an individual and population level, applying the approach to a test of English Language Proficiency that is scored using a four-dimensional IRT model.

Sunday, April 15, 2018
2:45-4:15pm, Plymouth, Individual Presentations, G7

Where Learning and Measurement Meet

Session Discussant: Kristen Huff, Curriculum Associates

The Effect of Peer Assessment on Learning: A Meta-Analysis

Hongli Li, Georgia State University; Yao Xiong, University of Pittsburgh; Charles Hunter, AdvancED; Xiuyan Guo, Emory and Henry College; Rurik Tywoniw, Georgia State University

Peer assessment encompasses processes whereby students evaluate or are evaluated by their peers. While there has been extensive research on peer rating accuracy, less attention has been paid to learning outcomes from this process. In this study, we conduct a meta-analysis to synthesize the effect of peer assessment on learning.

Teachers' Practices Related To Common Core State Standards-Aligned Assessments

Heather Buzick, Educational Testing Service; Cara Laitusis, Educational Testing Service; Teresa King, Educational Testing Service

Elementary and middle school ELA and mathematics teachers were surveyed about their instructional practices, test preparation strategies and test score use both before and after the introduction of CCSS-aligned assessments. This study documents evidence in support of the claim that CCSS-aligned assessments encourage better teacher practices.

Where are all these items coming from?

Amy Burkhardt, University of Colorado, Boulder; Derek Briggs, University of Colorado, Boulder

This paper focuses on the building blocks of interim assessments. It surveys the sources of item banks in popular assessment management systems used by large public school districts, and then focuses on two particular item banks being used within one large urban school district, exploring the evidence regarding item quality.

Factoring in effects of unified Opportunity-To-Learn framework on students' TIMSS mathematics achievement

Meiko Lin, Teachers College, Columbia University; Madhabi Chatterji, Teachers College, Columbia University

In evaluating ILSA results, an important contextual factor to consider is the Opportunity To Learn (OTL) for students in the subject area domains tested. The purpose of this study is to operationalize, validate and unpack OTL effects on Japanese students' mathematics achievement levels, building on a unified OTL framework.

Sunday, April 15, 2018
2:45-4:15pm, Manhattan, Individual Presentations, G8

Statistical Approaches to Improving Validity
Session Discussant: Ye Tong, Pearson

Adjusting Group Intercept Bias in Predictive Equations
Bruce Austin, Washington State University; Brian French, Washington State University

This study proposes and demonstrates an adjustment factor for predictive intercept bias found in common regression equations estimated using non-invariant group data. Results showed a large and consistent correction to predicted values from a common equation. This is especially beneficial when common predictive equations are used to make high-stakes decisions.

Accuracy of univariate and multivariate corrections for range restriction in practical applications
Tamar Kennet-Cohen, National Institute for Testing and Evaluation, Israel; Dvir Kleper, National Institute for Testing and Evaluation, Israel

We examined the performance of two corrections for range restriction –univariate and multivariate – by simulating a practical context where the weights of the predictors in the selection variable and the required statistics from the applicant pool are not necessarily known. The results confirmed the advantage of the multivariate correction.

A Statistical Procedure for Detecting Exactly or Nearly Matching Responses
Yi-Hsuan Lee, Educational Testing Service; Shelby Haberman, Edusoft

In many current tests, modern communication techniques can permit large numbers of examinees to share common response patterns to the entire test. This work presents a statistical procedure to identify examinees who exhibit unusual similarity in responses to the entire test in groups of any size greater than one.

Effects of Automated Rater Improvements on Test Equating Solutions
Michelle Boyer, University of Massachusetts and Data Recognition Corporation; Lisa Keller, University of Massachusetts; Richard Patz, University of California, Berkeley

Many studies have examined the quality of automated raters, but fewer have focused on potential effects related to psychometric properties of test scores. This study seeks to examine the comparability of test scores where advancing automated scoring technology is used to score short constructed response items in test equating scenarios.

Accounting for item alterations to link revisions of an ASL vocabulary test
James Davis, University of North Carolina at Greensboro; Jonathan Henner, University of North Carolina at Greensboro; Richard Luecht, University of North Carolina at Greensboro

Revisions of the ASLAI test of Synonyms were linked under the NEAT design within a Rasch measurement framework. The Mantel-Haenszel procedure was used to define an anchor set by identifying items that, despite minor alterations between versions, do not function differently between groups who took each version.

Sunday, April 15, 2018
2:45-4:15pm, Ambassador II, Coordinated Sessions, G9

Assessing mathematical thinking using learning progressions

Session Chair: Mark Wilson, University of California, Berkeley
Session Discussant: Richard Brown, National Math and Science Initiative
Session Discussant: James Pellegrino, University of Illinois at Chicago

Learning progressions are hypothesized structures that organize the topics of a mathematics curriculum into a representation of the ways that students increase in the sophistication of their mathematics thinking as they progress through the curriculum. It is crucial that (a) these hypothetical structures be empirically tested, and that (b) student locations along the progression can be well-assessed. This session includes four papers that illustrate one particular approach to measurement in the context of a learning progression, called the BEAR Assessment System (BAS). After an introductory presentation outlining the methods used, three papers discuss specific examples and report on (a) how well the learning progressions are measured, and whether the hypothesized structure is supported (or proposed to be modified), and (b) whether the supposed advantages of the learning progression approach seem to be borne out. The three papers are based on contexts that represent three different levels of complexity of the underlying learning progressions: (i) multidimensional, (ii) a two-dimensional example that has a common level at the top of both dimensions, and (iii) a seven-dimensional case that has cross-dimension requirements.

Supporting assessment in the context of a learning progression
Karen Draney, University of California, Berkeley; Mark Wilson, University of California, Berkeley

Assessing a learning progression in college-ready algebraic thinking
James Mason, University of California, Berkeley; Amy Arneson, University of California, Berkeley; Diah Wihardini, University of California, Berkeley; Mark Wilson, University of California, Berkeley

A complex learning progression structure for a critical statistical thinking construct
Amy Arneson, University of California, Berkeley; James Mason, University of California, Berkeley; Diah Wihardini, University of California, Berkeley; Mark Wilson, University of California, Berkeley

A learning structure involving requirements among dimensions for statistics and modeling
Mark Wilson, University of California, Berkeley; Perman Gochyyev, University of California, Berkeley; Richard Lehrer, Vanderbilt University

Sunday, April 15, 2018
4:35pm – 6:05pm, Majestic 2, Invited Session, H1

Award-Winning Research from the 2018 NCME Award Recipients

Session Moderator: Walter (Denny) Way, Pearson

2018 NCME Brenda H. Loyd Outstanding Dissertation Award
Statistical Learning Methods for Psychological Measurement
Yunxiao Chen, Emory University

2018 NCME Alicia Cascallar Award for an Outstanding Paper by an Early Career Scholar
Considering the Implications of Cumulative and Adjacent-Categories Models for Raters: An Illustration using Mokken Scale Analysis
Stefanie Wind, The University of Alabama

2018 NCME Jason Millman Promising Measurement Scholar Award
The reality of subscore reporting: balancing measurement and policy perspectives
Richard Feinberg, National Board of Medical Examiners

2018 NCME Annual Award
The Stanford Educational Data Archive (SEDA): Using measurement methods to make public test score data useful
Sean Reardon, Stanford University; Andrew Ho, Harvard University; Benjamin Shear, University of Colorado Boulder; Erin Fahle, Stanford University; Demetra Kalogrides, Stanford University; Ken Shores, University of Pennsylvania; Katherine Castellano, Educational Testing Service

2018 NCME Bradley Hanson Award for Contributions to Educational Measurement
Data Forensics Analysis: Continuing the Legacy of Bradley Hanson
Sandip Sinharay, Educational Testing Service

2018 NCME Career Award
Brian Clauser, National Board of Medical Examiners (to present at the 2019 NCME Annual Meeting)

Sunday, April 15, 2018
4:35-6:05pm, Ambassador III, Coordinated Sessions, H2

The Big Five (Sources of Validity Evidence): Illustrations of Validation Practices

Session Chair: Stephen Sireci, University of Massachusetts Amherst
Session Discussant: Michael Kane, Educational Testing Service

For over 60 years, the National Council on Measurement in Education has been a full partner in the development of the *Standards for Educational and Psychological Testing*. The past two versions of the *Standards* described five "sources of evidence that might be used in evaluating the validity of a proposed interpretation of test scores for a particular use" (AERA, APA, & NCME, 2014, p. 11). Although the *Standards* are well-known, the five sources are not; and validation practices and measurement textbooks still cling to outdated terminology and validation approaches. In this symposium, we bring together an experienced panel of measurement experts to illustrate theories and methods associated with each source of validity evidence and provide examples of applied, 21st-century validity studies. These presentations will be followed by a discussion of how evidence provided from these five sources can be synthesized into a comprehensive validity argument.

Evaluating Criteria for Validity Evidence Based on Test Content
Stephen Sireci, University of Massachusetts Amherst; Ella Banda, University of Massachusetts Amherst; Gabriel Rodriguez, University of Massachusetts Amherst; April Zenisky, University of Massachusetts Amherst; Hwanggyu Lim, University of Massachusetts Amherst

Validity Evidence Based on Response Processes
Susan Embretson, Georgia Institute of Technology

Evaluating Internal Structure Using Factor Analytic Techniques
Craig Wells, University of Massachusetts Amherst; Francis O'Donnell, University of Massachusetts Amherst

Assessments of College Readiness: Validity Evidence Based on Relations to Other Variables
Krista Mattern, ACT, Inc.

Validity Evidence Based on Consequences of Assessment and Accountability Programs
Suzanne Lane, University of Pittsburgh; Danielle Niepokoj, University of Pittsburgh

Sunday, April 15, 2018
4:35-6:05pm, Melville, Coordinated Sessions, H3

Dimensionality as it Relates to Primary Latent Factors, Sub-scores, and Item Parcels

Session Chair: Ernest Davenport, University of Minnesota
Session Discussant: Steven Culpepper, University of Illinois at Urbana-Champaign

This symposium explores dimensionality from several perspectives: latent factors, subscores, and item parcels; covering the literature on item response theory, factor analysis, subscores, and item parcels. The qualities of a composite may appear different coming from the combination of subscores measuring different latent entities, than from IRT where one assumes all contributing items are unidimensional. The definition of dimensionality provided by Cronbach (1951) unifies the meaning of a composite from these disparate perspectives. Dimensionality is not a binary concept (unidimensional versus multidimensional). Using the variance accounted for by the first principal factor of the items to index dimensionality is appropriate as this variance measures the vector in the space of the items with maximal relationship to whatever the items share in common. Given our view of dimensionality, we show fairly common conditions which allow a composite to be interpreted similarly regardless of how it is obtained. The utility of a composite does not depend solely on whether the assessment items are unidimensional at the primary level. Inter-relationships of the items on multiple dimensions and/or correlations of the separate dimensions can lead to general factors whose composites are meaningful.

Dimensionality and the Meaning of Composites
Ernest Davenport, University of Minnesota

Dimensionality as it Relates to Factor Analysis and Item Response Theory
Cengiz Zopluoglu, University of Miami

Dimensionality for Composites from SubScores from Different Latent Entities
Mark Davison, University of Minnesota

Dimensionality and Item Parcels
Youngsoon Kang, University of Minnesota

Sunday, April 15, 2018
4:35-6:05pm, Majestic I, Invited Session, H4

New Developments in the Assessment Practice at the National Center for Assessment
Session Moderator: Linda Cook, Educational Testing Service
Discussant: Kurt Geisinger, University of Nebraska/Buros Center for Testing
Discussant: William G. Harris, Association of Tet Publishers

Multi-stage adaptive testing, D-scoring implementation, and Tablet IQ battery are some of the new developments at the National Center for Assessment (NCA). An approach to test scoring and equating, referred to as 'delta-scoring' (or '*D*-scoring'), that is under implementation in the assessment practice of the National Center for Assessment (NCA) will be presented. For a test with binary items, the *D*-score of an examinee is based on his/her response vector on the items weighted by the expected difficulty. D-score is scaled within the range of 0 to 1 to reflect what proportion of the ability required for a total success on the test is demonstrated by the examinee.

Among the large-scale testing developments, the merits of the adaptive testing have been employed but the usual inherent problems of miss-representation of the content or test dimensions have been deferred. A three-stage and four levels adaptive model has been employed in the development of a high-stake university admission test.

Lately, a new generation of IQ tests has been developed at the NCA with many advantages over the kit oriented tests. This new test battery is made as an interactive session on the I-Pad minimizing the clinician effect in test administration, scoring, and reporting. Thus, the test is belt as an adaptive one with 48 subtests. The test battery is based on two theories; namely, C-H-C theory of intelligence and PASS theory.

The National Center for Assessment: Testing Tools and Multistage Testing Application
Faisal Al Saud, National Center for Assessment

A New Method of Test Scoring and Equating (D-scoring) and its Application in the Assessment Practice at the NCA
Dimiter Dimitrov, National Center for Assessment, Saudi Arabia and George Mason University

A New Generation of Individual IQ Test: The Adaptive I-Pad "Qiyas Battery for Intelligence"
Khaleel Al Harbi, National Center for Assessment; Elena Grigorenko, University of Houston; Abdullah Al Qataee, National Center for Assessment

Sunday, April 15, 2018
4:35-6:05pm, Gershwin II, Individual Presentations, H5

Diving into Data with Response Process Research

Session Discussant: Michael Kane, Katie McClarty, Questar

Using Process Data to Explain Group Differences in Complex Problem Solving

Beate Eichmann, German Institute for International Educational Research; Frank Goldhammer, German Institute for International Educational Research; Samuel Greiff, University of Luxembourg; Liene Pucite, Goethe University Frankfurt; Johannes Naumann, Goethe University Frankfurt

We used computer-generated log data to extract behavioral indicators and investigate their adequacy for explaining performance differences between groups in complex problem solving. Our results indicate that exploration mediates the effect of gender but not the effect of migration background and should be encouraged in education to diminish performance differences.

Analyzing the Process of Clinical Diagnosis with Educational Data Mining

Feiming Li, Zhejiang university of Technology

Computer-based environment enables assessment of clinical diagnostic reasoning and data gathering skills in more dynamic and interactive but nonintrusive and less expensive way. This study analyzed those skills underlying the final diagnosis decision with educational data mining methods based on the log files recording examinees' processing actions.

Invariance of the Response Process between Modes and Gender in Reading Assessment

Ulf Kroehne, DIPF (German Institute for International Educational Research); Frank Goldhammer, DIPF (German Institute for International Educational Research); Carolin Hahnel, DIPF (German Institute for International Educational Research)

Reading scores can differ between administration conditions (computer vs. paper) and groups (e.g., gender). In this paper we use log-data and a bivariate generalized linear IRT model (Molenaar et al., 2015) to investigate the response process by testing invariance of the relationship between speed and ability across modes and gender.

Hints, Multiple Attempts, and Learning Outcomes in Computer-based Formative Assessment

Jinnie Choi, Pearson; Mikolaj Bogucki, Pearson

This study analyzed data from a computer-based formative assessment system in which instructors can control the settings around the number of question attempts and how student opening of hints are scored. Evidence showed that more formative settings were associated with higher persistence and conscientiousness, but lower performance.

Sunday, April 15, 2018
4:35-6:05pm, Belasco, Individual Presentations, H6

Modeling Response Times

Session Discussant: Patrick Kyllonen, Educational Testing Service

Regression Tree Modeling for the Prediction of Response Time on Pretest Items
Qiongqiong Liu, National Board of Osteopathic Medical Examiners; Isaac Li, National Board of Osteopathic Medical Examiners; Yi Wang, National Board of Osteopathic Medical Examiners; Edward Tsai, National Board of Osteopathic Medical Examiners

This study uses regression tree models to predict the response time of pretest items in a licensure examination based on item characteristics prior to test administration, which can enhance the parallelism of assembled examination forms. This technique is compared to the linear regression approach in prediction accuracy.

Measuring English language proficiency across subgroups with response time data
Hanwook (Henry) Yoo, Educational Testing Service; Venessa Manna, Educational Testing Service

This study evaluates potential benefits of using response time (RT) data as it pertains to interpretation of score differences in the context of English language proficiency. Both test- and item-level RT are analyzed to validate subgroup difference categorized by background characteristics (gender, native language, and years spent studying English).

Weighted Likelihood Estimation Method for Response and Response Time
Anqi Li, University of Illinois at Urbana-Champaign; Hua-Hua Chang, University of Illinois at Urbana-Champaign

This study extends the weighted likelihood estimation (WLE) method for both response accuracy and response time. First, WLE method is derived for the hierarchical framework modeling both response accuracy and response time. Then the extended WLE method is compared with maximum likelihood estimation (MLE) and Bayesian estimation methods.

A Hierarchical Framework for Response Times and Signal Detection Theory
William Muntean, Pearson Vue; Joe Betts, Pearson Vue; Shu-chuan Kao, Pearson Vue; Ada Woo, National Council of State Boards of Nursing

Signal detection theory is useful in determining the sources of responding error—either in responding bias (over/under selecting responses) or sensitivity. The current work extends signal detection theory by incorporating response times under a unified hierarchical model to provide a richer and more holistic view of multiple response data.

Sunday, April 15, 2018
4:35-6:05pm, Plymouth, Individual Presentations, H7

Scoring Simulations, Performance Tasks, and Polytomous Items
Session Discussant: Howard Everson, SRI International

Exploring Alternative Scoring Methods for Large-Scale Computer-Based Case Simulations
Dandan Liao, University of Maryland, College Park; Matthias von Davier, National Board of Medical Examiners; Jonathan Rubright, National Board of Medical Examiners

This study focuses on computer-based case simulations in a large-scale licensure exam. Both exploratory data analysis and confirmatory psychometric modeling approaches were utilized to explore structured process data collected during the exam. Results suggest that data-driven methods can produce more reliable ability estimates than rule-based scoring methods.

Evaluating a complex technology-based assessment (TBA) to measure teachers' instructional performance
Olga Zlatkin-Troitschanskaia, Johannes Gutenberg-University Mainz, Department of Business and Economics Education; Christiane Kuhn, Johannes Gutenberg-University Mainz, Department of Business and Economics Education; Jacqueline Leighton, University of Alberta, Department of Educational Psychology; Hannes Saas, Johannes Gutenberg-University Mainz, Department of Business and Economics Education; Sebastian Brückner, Johannes Gutenberg-University Mainz, Department of Business and Economics Education

Teachers´ instructional performance was modeled and measured as a complex interaction of both domain-specific knowledge and generic skills using a newly developed complex technology-based assessment (TBA). We present validity evidence from cognitive interviews with 42 teachers (novice and experts) and reveal that TBAs can enhance instructional performance.

Score Resolution for Medical Certification Portfolio Exams
Lisa Reyes, Measurement Incorporated

Using many-facet Rasch measurement to identify score profiles exhibiting rater disagreement in a two-rater scoring system, this study investigates how incorporating a third set of ratings to resolve rater disagreement impacted pass-fail outcomes on a portfolio certifying exam. Results and implications for portfolio exam score resolution methods will be discussed.

Investigating Rater Effects in International Large-Scale Assessments
Hyo Jeong Shin, Educational Testing Service; Matthias von Davier, National Board of Medical Examiners; Kentaro Yamamoto, Educational Testing Service

The present research investigates rater effects in international large-scale assessments to evaluate the validity of assigned scores by human raters and to improve the precision of group-level reporting scores. We present a multilevel item response model for this situation and apply the model to PISA data collected in 2015.

Evaluating the Utility of the Diagnostic Rating System for Performance Assessment
Nicholas Curtis, James Madison University; Allison Ames, James Madison University

A new scoring method for performance assessment, grounded in logic-based decision mapping, has shown much promise in initial testing. The current, in-depth study evaluates the validity of the interpretations of scores, cognitive load, rater effects, and response time compared to traditional rubric methods.

Sunday, April 15, 2018
4:35-6:05pm, Manhattan, Individual Presentations, H8

IRT with Non-traditional Constructs

Session Discussant: James Roberts, Georgia Institute of Technology

IRT Mixture Model for Rating Scale Confusion Associated with Negatively Worded Items
Daniel Bolt, University of Wisconsin, Madison; Yang Wang, Education Analytics Inc.; Robert Meyer, Education Analytics Inc.; Andrew Rice, Education Analytics Inc.

We illustrate application of a mixture IRT model to address respondent confusion related to the negative wording of items in a social-emotional learning assessment. Application of the model (1) confirms confusion primary occurs at lower grade levels (3rd-5th), and (2) corrects for bias in evaluating the psychometric performance of scales.

Psychometric properties of a measure of students' interpersonal and intrapersonal skills
Megan Kuhfeld, NWEA

Non-cognitive domains have been found to be predictive of key adult outcomes, but these domains have not been well-validated compared with cognitive domains. This study uses high school student surveys to examine the psychometric properties of a set of intrapersonal/interpersonal competencies.

Response Processes in Noncognitive Measures: Validity Evidence from Explanatory Item Response Modeling
Michael Rodriguez, University of Minnesota; Okan Bulut, University of Alberta; Julio Cabrera, University of Minnesota; Jose Palma, University of Minnesota

Consistent with improving the positive impact of assessment on teaching and learning, we explore score interpretation validation for a noncognitive measure of *Social Competence*, using a partial-credit explanatory item response model. We find item and person characteristics interact in significant ways, influencing test-taker response processes and potentially influencing score interpretation.

Adopting a Process Perspective of Collaborative Problem Solving
Sandra Milligan, University of Melbourne; Mark Wilson, The University of California, Berkeley; Patrick Griffin, University of Melbourne; Claire Scoular, University of Melbourne; Daniel Jimenez Barrios, University of Melbourne; BM Monjurul Alom, University of Melbourne; Nafisa Awwal, University of Melbourne; Zhonghua Zhang, University of Melbourne

The authors proposed a new Collaborative Problem Solving (CPS) Process Framework. This framework presents CPS as an amalgamation of collaboration and problem solving frames. Auto-scored indicators and a multifaceted scoring algorithm have been developed for this framework. Empirical data will be analysed to explore the validity evidence for this framework.

Sunday, April 15, 2018
4:35-6:05pm, Ambassador II, Coordinated Sessions, H9

Communicating Complex Psychometric Information to Teachers, Parents, and Other Less Technical Audiences

Session Chair: Karen Draney, University of California, Berkeley

This symposium presents a series of ideas for communicating both psychometric methods, and the results of these, to audiences with little or no psychometric training. Such audiences may include the typical school audiences of teachers, students and parents, but also may include district and state personnel who may need to understand the results of an assessment program. The symposium begins with a description of a philosophical and methodological approach to communicating complex psychometric information. The ensuing presentations demonstrate examples of this approach: methods for including early childhood educators in an empirically-based system for setting Kindergarten readiness standards, and score reports designed for (a) parents and teachers of children in infant-toddler, preschool, and Kindergarten programs and (b) legislators and personnel in the Department of Education. The ideas and approaches in this symposium demonstrate that good communication of complex psychometric information is an important aspect of assessment systems that serve the needs of multiple stakeholders and ultimately support teaching and learning.

Public Understanding of Educational Measurement: Communicating Psychometric Methods And Outcomes
Perman Gochyyev, University of California, Berkeley; Mark Wilson, University of California, Berkeley

The Criterion-Zone Boundary Approach for Communicating Student Progress
Karen Draney, University of California, Berkeley; Linda Morell, University of California, Berkeley; Kerry Kriener-Althen, WestEd

High Quality Score Reports from a Multidimensional Assessment of Early Childhood Development
Joshua Sussman, University of California, Berkeley; Rebecca Freund, University of California, Berkeley; Leah Feuerstahler, University of California, Berkeley

Sunday, April 15, 2018
4:35-6:05pm, Gershwin I, Individual Presentations, H10

Electronic Board Session 3

Electronic Board #1

Linkability Analysis Focused on Reliability of Linked Scores

Yoshikazu Sato, Kyushu University; Tadashi Shibayama, Tohoku University

Under a single group design, we have developed new formulas for the root mean square error (*RMSE*) of measurement and the reliability coefficient, as basic indices for the reliability of linked scores; and have shown that these indices can be applied to linkability analysis.

Electronic Board #2

Impact of Nonignorable Missing Data on the Performance of Person Fit Statistic

Xin Qiao, University of Maryland College Park

Current study was to explore the effects of nonignorable missing on the person fit statistic (*lz*) with dichotomous items in the context of cheating behavior. Three missing data treatment (MDT) methods were investigated under various testing conditions through ROC analysis. Practical suggestions were given on the choice of MDT methods.

Electronic Board #3

Measuring Instructional Sensitivity: The Role of Covariance Structures on the Group Level

Alexander Naumann, German Institute for International Educational Research (DIPF); Johannes Hartig, German Institute for International Educational Research (DIPF); Jan Hochweber, University of Teacher Education St. Gallen (PHSG)

Valid inferences on schooling and teaching drawn from students' test scores require that tests and items are instructionally sensitive. However, there is little knowledge on the relationship of test sensitivity and item sensitivity. Thus, the present study aims at investigating how item sensitivity relates to test sensitivity.

Electronic Board #4

Subgroup Identification Using Regression Trees in a Scenario-based Writing Assessment

Yi Cao, Educational Testing Service; Jianshen Chen, Educational Testing Service; Mo Zhang, Educational Testing Service; Paul Deane, Educational Testing Service

Scenario-based assessments intend to better inform and support teaching and learning. This study used regression trees to identify heterogeneous subgroups on performance differences between writing assessments with and without scenario-based lead-in tasks via writing-process features and background covariates. Results can inform personalized training and the development of customized assessments.

Electronic Board #5

Is a Wald-type Test approach useful in selection of 2PL items?

Hirotaka Fukuhara, Pearson; Insu Paek, Florida State University

This study explores a utility of a Wald-type test in making a decision on an IRT model between 2PL and 3PL for dichotomously scored items. Also, the performance of the Wald-type test is compared with the likelihood ratio test approach with a mixture chi-square distribution via simulation.

Electronic Board #6

Evaluating simulated rater uncertainty in Angoff and modified Angoff standard setting results

Kirk Becker, Pearson; James Masters, Pearson; Haiqin Chen, American Dental Association

Reckase (2006) used the concept of a panelist's intended cut score with simulated standard setting data to evaluate bookmark and modified Angoff standard setting methods. This paper will incorporate the concept of intended cut score to explore the effect of error or bias on Angoff yes/no and modified Angoff ratings.

Electronic Board #7

Comparing the Normalized and 2PL IRT Scoring Methods on Multi-session Exams

Aolin Xie, Prometric. Inc; Ting-Wei Chiu, Prometric.Inc; Keyu Chen, Prometric. Inc; Greg Camilli, Rutgers, The State University of New Jersey

This study compared the candidates' scores based on the normalized model and the 2-parameter Item Response Theory (IRT) model. Data were simulated using item parameters obtained from a multi-session exam. Candidates' calculated scores, rankings, pass/fail statues and score ties from the two models were compared with their true values.

Electronic Board #8

Highlighting Response Processes within Validation Efforts for an Inference-based Multiple-Choice Assessment

Tia Fechter, Pacific Metrics; Jennifer Cromley, University of Illinois at Urbana-Champaign; Martin Van Boekel, University of Minnesota-Twin Cities; Ting Dai, Temple University; Frank Nelson, Temple University; Aygul Dane, University of Illinois at Urbana-Champaign

The emphasis of this presentation is on the efficacy of using cognitive interviews and a refined coding scheme to highlight different cognitive processes engaged by examinees who respond correctly and incorrectly to inference-based multiple-choice questions as a way to evaluate response processes within a larger assessment validity framework.

Electronic Board #9

An Investigation of the Dimensionality of a Large Scale Hybrid Assessment

Lei Wan, The College Board; HyunJoo Jung, University of Massachusetts Amherst; Pamela Kaliski, The College Board

In this study, the dimensionality of a large scale hybrid assessment, comprised of two through-course performance tasks and one end-of-course exam, was examined. To investigate the structure of the scores from the assessment, response data from 18,067 examinees were used in the structural equation modeling and item response theory analyses.

Electronic Board #10

Classification Consistency and Accuracy in Multiple Measures using IRT

Seohee Park, University of Iowa; Kyung Yong Kim, University of North Carolina at Greensboro; Timothy Ansley, University of Iowa; Ariel Aloe, University of Iowa

The purpose of this study is to extend the existing IRT-recursive based classification consistency and accuracy indices into multiple-measure situations. Three decision-making rules, which are the complementary, conjunctive and compensatory rules, and pairwise combinations of the three rules will be presented.

Electronic Board #11

Investigating Grade-Level Differential Item Functioning in CAT via Bayesian Inference

Johnny Denbleyker, Houghton Mifflin Harcourt

A large-scale GK-12 adaptive assessment having 4,110 unique Math items available across grades 2-9 was utilized to compare item difficulty parameters. By having items administered across multiple grade-levels and calibrated via a fixed-theta anchored design, grade-level DIF questions are investigated via Bayesian methods with multiple perspectives.

Electronic Board #12

Effects of Low Category Response Rates on Common Polytomous IRT Item Calibrations

Ki Cole, Oklahoma State University; Ki Cole, Oklahoma State University; Insu Paek, Florida State University

When an ordered polytomous response scale is used, proportions of responses in the end categories tend to have low response rates. The purpose of this simulation study was to investigate and compare the impacts of low-response categories on item calibrations with the graded response and generalized partial credit models.

Electronic Board #13

Impact of Semantic Similarity to Training Responses on Automated Scoring Accuracy

Richard Meisner, ACT, Inc.

Using semantic word vectors, similarities between examinees' written short responses and exemplar training responses were quantified, and the similarity values were used as predictive inputs in a machine learning model for automated scoring in the attempt to improve accuracy. Scoring improvements were noted for many content areas.

Electronic Board #14

Model Comparison for a Testlet-based English Language Assessment of Young Students

Tongyun Li, Educational Testing Service; Jiyun Zu, Educational Testing Service

The proposed study is an investigation of the testlet effect in a large-scale English assessment of young students. Standard IRT, testlet and bifactor models are respectively fit to the three sections of the assessment. Results will illustrate the trade-off between model parsimony and modeling local dependence in operational settings.

Electronic Board #15

Building a Cloud-based CAT System for a Large-Scale Language Proficiency Test

Jing Yang, Northeast Normal University; Liwen Huang, University of Illinois at Urbana-Champaign; Leanne Zeng, University of Illinois at Urbana-Champaign; Hua-Hua Chang, University of Illinois at Urbana-Champaign

This study compared two item selection methods, *b*-matching with ascending *a*-stratification and Maximum Priority Index (MPI) with ascending *a*-stratification, using a CAT item pool from a large-scale language proficiency test, and found that *b*-matching optimizes item pool usage while retaining good estimate accuracy and exposure control, compare to MPI.

Electronic Board #16

Readability Measures for Multiple-Choice and Innovative Items

Natalie Jorion, PearsonVUE; William Muntean, PearsonVUE; Joe Betts, PearsonVUE; Doyoung Kim, National Council of State Boards of Nursing; Ada Woo, National Council of State Boards of Nursing; Philip Dickison, National Council of State Boards of Nursing

Evaluating readability is important when determining the dimensionality of an assessment, especially when one dimension has disproportionally lengthy stimuli (e.g., charts, tables, narratives, etc.). The current study explores multiple methods for determining the contributions (and lack thereof) of readability in detecting an artificial dimension in innovative items.

Electronic Board #17

Investigating Score Comparability of Computer Adaptive and Linear Testing

Tianli Li, ACT Inc.; Jie Li, ACT Inc.

This study evaluates the performance of four scaling methods that produce comparable scale scores for CAT with its corresponding linear tests under the situation that the CAT and the linear tests are scored using theta estimates and number-correct scores, respectively.

Electronic Board #18

Mode Transitions of Group Score with Non-Random Samples

Lingyun Gao, Measured Progress, Inc.; Steve Wise, NWEA; Quinn Lathrop, Pearson

This study examines the impact of mode delivery change on group scores for a large-scale assessment. To adjust group score differences across modes, a conditioning variable representing mode effects was added to the scoring model. Post-stratification weights were also applied to improve data representativeness.

Electronic Board #19

LOO and WAIC as IRT Model Selection Methods with Mixed-Format Tests

Yong Luo, National Center for Assessment, Saudi Arabia

LOO and WAIC are two fully Bayesian model selection methods that have shown promise for both dichotomous and polytomous data. This study investigates their performances as model selection methods with mixed-format tests in comparison with common model selection criteria such as LRT, AIC, BIC, and DIC.

Monday, April 16, 2018
8:15-10:15am, Broadway I, Coordinated Sessions, I2

Mapping state proficiency standards to the NAEP scale: New methods, new results

Session Chair: Benjamin Shear, University of Colorado Boulder
Session Chair: Sean Reardon, Stanford University
Session Chair: Joe Willhoft, National Assessment Governing Board
Session Discussant: David Thissen, University of North Carolina

Since the No Child Left Behind Act of 2001, federal accountability policies have required each state to set its own "challenging academic standards" in mathematics, reading or language arts, and science (ESSA, 2015; NCLB, 2002). Variability in standards across states is an educationally and politically relevant concern—the public should know whether educators and policymakers are holding students to higher standards in, for example, Massachusetts than Mississippi. However, methods for comparing performance standards across different tests are not straightforward and rest on particular assumptions.

This symposium begins with a critical review of two different approaches for comparing state standards, both using the National Assessment of Educational Progress (NAEP). The first approach uses equipercentile linking; the second uses a sequential application of heteroskedastic probit modeling and linear linking. The symposium also offers new results through 2015, including a summary of how state standards have risen over time and converged in variability through the so-called "common core" era. Finally, there will be discussion of the role of the NAEP as the de facto basis and benchmark for comparing standards across states, from the perspective of a panelist involved in the recent National Academy of Sciences review of NAEP achievement levels.

Mapping state proficiency standards onto the NAEP scales
Taslima Rahman, National Center for Educational Statistics; Victor Bandeira de Mello, American Institutes for Research

Mapping proficiency standards using heteroskedastic ordered probit models and NAEP-based linear linking
Benjamin Shear, University of Colorado Boulder; Sean Reardon, Stanford University; Erin Fahle, Stanford University; Andrew Ho, Harvard University

Rise and convergence of state proficiency standards in the Common Core era
Andrew Ho, Harvard University; Benjamin Shear, University of Colorado Boulder; sean reardon, Stanford University; Erin Fahle, Stanford University

Intended uses of NAEP proficiency standards: Perspectives from the National Academies Evaluation
Laura Hamilton, RAND Education

Monday, April 16, 2018
8:15-10:15am, Broadway II, Coordinated Sessions, I3

Emergent Themes from the Development of NGSS-Aligned Summative Science Assessments

Session Chair: Lei Liu, Educational Testing Service
Session Chair: Michelle Center, California Department of Education
Session Discussant: James Pelligrino, University of Illinois at Chicago

The Next Generation Science Standards, describes a new vision for science learning, teaching, and assessment whereby science proficiency is an integrated understanding of three dimensions – disciplinary core ideas, science and engineering practices, and crosscutting concepts – in the form of *performance expectations*. To design NGSS-aligned student assessment, it is expected that the assessment should provide useful information related to the multidimensional nature about student science learning to all stakeholders including students, teachers, administrators, policy makers, and the public. Essentially assessments aligned to the NGSS should generate interpretable data that yield a positive impact on science teaching and learning, consistent with the theme of the 2018 NCME conference. Given the breadth and depth of material covered in the NGSS, the standards present new challenges to assessment designers and call for new approaches to overcoming obstacles in various aspects of the development process, including assessment designing, scoring, reporting, and accessibility. As one of the NGSS lead states, California became one of the first to embrace these challenges as part of a joint venture between the California Department of Education and the Educational Testing Service. This coordinated session reports on accumulated findings from both research and field experiences during the development process.

Using the ECD Process to Design NGSS-Aligned Items
Lei Liu, Educational Testing Service; Gary Weiser, Educational Testing Service; Janet Koster van Groos, Educational Testing Service; Oliver Islambouli, Educational Testing Service

Challenges and Proposed Solutions to Score Reporting
Longjuan Liang, Educational Testing Service; Katherine Castellano, Educational Testing Service

Accessibility Challenges and Opportunities
Danielle Guzman-Orth, Educational Testing Service; Teresa King, Educational Testing Service; Cara Laitusis, Educational Testing Service; Cary Supalo, Educational Testing Service

Monday, April 16, 2018
8:15-10:15am, Broadway III, Coordinated Sessions, I4

Diagnosis and Feedback in Learning and Assessment Systems

Session Chair: Jimmy de la Torre, The University of Hong Kong
Session Chair: Alina von Davier, ACTNext
Session Discussant: Jacqueline Leighton, University of Alberta

College admission tests are comprehensive educational assessments that high school students take more than once to improve their chances of admission into the college of their choice. In addition, students prepare for the test by taking online test preparation programs. To help students master the skills required in college, these learning and assessment systems need to provide actionable diagnosis and feedback. A well-designed online preparation system should be capable not only of providing students diagnostic feedback about skills at a granular level, but also access to remedial and learning resources. In this session, we illustrate a comprehensive approach to diagnosis and feedback based on cognitive diagnostic and learning models. The four papers in this coordinated session that look at different components of learning and assessment systems cover attribute and Q-matrix development, as well as model selection for extant ACT tests; supplementing information obtained from ACT tests by administering additional items using a computerized adaptive testing approach; modeling student's test preparation and performance using knowledge tracing techniques; and developing a more reliable online rating system. This coordinated session aims to provide meaningful thoughts on possible solutions from different methodological perspectives on challenges encountered when building an online learning and preparation system.

Testing for learning: Actionable instructional feedback based on the ACT Test
Alina von Davier, ACTNext; Pravin Chopade, ACTNext; Pamela Paek, ACT; Jimmy de la Torre, The University of Hong Kong; Melanie Rainbow-Harel, ACT; David Carmody, ACT

What does it take to provide accurate feedback? A simulation study
Yan Sun, Rutgers University; Pravin Chopade, ACTNext; Alina von Davier, ACTNext; Jimmy de la Torre, The University of Hong Kong

Modeling skill evidence from test preparation learning behaviors
Steve Polyak, ACTNext; Michael Yudelson, ACTNext

Urnings: A rating system for learning analytics
Gunter Maris, ACTNext; Han van der Maas, University of Amsterdam; Maria Bolsinova, University of Amsterdam; Abe Hofman, University of Amsterdam; Matthieu Brinkhuis, Utrecht University; Benjamin Deonovic, ACTNext; Jesse Koops, Cito

Monday, April 16, 2018
8:15-10:15am, Ambassador III, Individual Presentations, I5

Bayesian Applications

Session Discussant: Michael Edwards, Arizona State University

A Bayesian Synthesis Approach to Data Fusion Using Augmented Data-Dependent Priors

Katerina Marcoulides, University of Florida

Data fusion involves merging datasets sharing some common variables to create a new dataset permitting more flexible analyses than when examining the data separately. This study investigates a new data fusion approach called Bayesian Synthesis. Results illustrate the effectiveness and practical utility of the approach for conducting assessment analyses.

Bayesian Expectation-Maximization-Maximization-Maximization Algorithm for the 4PLM

Ci Zhang, University of Illinois at Urbana-Champaign; Shaoyang Guo, University of Illinois at Urbana-Champaign; Chanjin Zheng, Jiangxi Normal University

There is a renewed interest in the 4PLM, but a user-friendly calibration method constitutes a major barrier to its widespread application. Bayesian Expectation-Maximization-Maximization-Maximization (BEMMM) is proposed based on a latent-mixture-modeling reformulation. The results indicated that BEMMM is as accurate as MCMC and as fast as the Bayesian modal estimation.

Modeling Response Bias in Polytomous Data Using the Bayesian Approach

Jiaqi Zhang, University of Cincinnati; Lihshing Wang, University of Cincinnati

Since response bias exists in polytomous data, we fit a 3PL model under framework of GRM and compare the result with the 2PL-GRM using a simulation study and a real data analysis in Bayesian approach. 2PL-GRM results in biased estimations, and 3PL-GRM can account for sources of response biases.

Modeling the Intertrait Correlation in Bayesian MIRT Models Using Separation Strategy

Meng-I Chang, Southern Illinois Uiniversity Carbondale; Yanyan Sheng, Southern Illinois University Carbondale

This study applied a Separation Strategy via the use of the LKJ prior in modeling the covariance matrix of a Bayesian multi-unidimensional IRT model, and further compared it with the conventional approach, i.e., the inverse Wishart prior. Results showed that the former performs better than the latter under most conditions.

Analysis of Incomplete Ordinal Data in Structural Equation Modeling Using Bayes Estimator

Yan Xia, Arizona State University; Yi Zheng, Arizona State University

We compared the Bayes estimator with WLSMV and ULSMV for ordinal incomplete data in the analysis of structural equation modeling. We expect Bayes estimator outperforms WLSMV and ULSMV when missing data exist, because the former simulates the underlying continuous variables conditioning on the other information in the MCMC process.

Bayesian Approach to Embedded Item Calibration

Bingnan Jiang, ACT. Inc.

Monday, April 16, 2018
8:15-10:15am, Majestic I, Individual Presentations, I6

Automated Scoring

Session Discussant: Mark Shermis, University of Houston

Adaptive scoring of constructed response: Issues and possibilities

Isaac Bejar, Educational Testing Service; Michael Kane, Educational Testing Service; Steven Holtzman, Educational Testing Service; Kevin Larkin, Educational Testing Service

Although human scoring is now conducted online, the process still reflects its paper-and-pencil origin. We describe the concept of adaptive scoring whereby the responses from students would be scored more or less precisely depending on how close they are to cutscores. We discuss psychometric, policy and implementation issues.

Meta-Analytic Methods for Assessing Agreement in Automated Essay Scoring Systems

Betsy Becker, Florida State University; JiYeo Yun, Florida State University

We use meta-analytic methods to evaluate distributions of indices of agreement between human and automated essay scores. We compute the likelihood of finding agreement indices in particular regions of interest, according to preset criteria for "good enough" agreement. The techniques are illustrated using meta-analytic data on writing assessments.

Creativity on ICE: Evaluating Contribution of Essay Features to Automated Creativity Predictions

Brad Bolender, ACT; Dan Shaw, ACT; Richard Meisner, ACT

In this study, Individual Conditional Expectation (ICE) plots were used to analyze the contributions of features to creativity score predictions in an automated scoring model trained on a sample of 3,603 essays. Understanding the contributions of these features could aid the development of automated measures of creativity in writing.

Industry Standards for an Emerging Technology: Automated Scoring

Lisa Haisfield, ACT; Erin Yao, ACT

Automated Scoring (AS), used to score constructed response and essay items, has become more widely integrated into education in recent years. However, skepticism still exists around this technology. This presentation will review and highlight AS industry standards that are intended to produce confidence for assessment stakeholders using using AS technology.

A generalizability theory approach to the PEG automated essay scoring system

Dandan Chen, University of Delaware; Joshua Wilson, University of Delaware; Michael Hebert, University of Nebraska–Lincoln; Micheal Sandbank, University of Texas at Austin

We investigated the number of prompts needed to obtain a generalizable estimate of elementary students' writing assessed by the PEG Writing® automated essay scoring system. To reach generalizability of .90, struggling writers needed one additional prompt per genre than non-struggling writers. Implications for feasible evaluation of writing ability are considered.

Monday, April 16, 2018
8:15-10:15am, Plymouth, Individual Presentations, I7

Advances in Communicating Results

Session Discussant: Priya Kannan, Educational Testing Service

Low Scoring Examinees Have More Variable Score Profiles: More than Just Error?
Mark Raymond, National Board of Medical Examiners

A common approach to evaluating subscores is to inspect subtest correlations for subgroups of examinees. The present study evaluates an index of score profile variability, and illustrates how the binomial error model and multivariate generalizability theory can be used to differentiate signal from noise in subgroup score profiles.

Teacher Use of Score Reports for Instructional Decision-Making
Amy Clark, University of Kansas; Meagan Karvonen, University of Kansas; Russell Swinburne Romine, University of Kansas; Neal Kingston, University of Kansas

This presentation describes results from teacher focus groups on the use of alternate assessment (AA-AAS) score reports for instructional planning. Summative score reports delivered from the 2016-2017 academic year serve as the basis for discussion on how results are used during the fall of 2018 to guide instructional decision-making.

Exploring and Visualizing School Achievement and School Effects
Daniel Anderson, University of Oregon; Joseph Stevens, University of Oregon

In this presentation we discuss methods for visualizing differences in school effect estimates. Specifically, we discuss joy plots – an alternative to standard caterpillar plots – as well as geo-spatial mappings to explore the relation between school effect estimates and the demographic characteristics of the school's surrounding area.

Investigating Score Reports for Universal Screeners: Do they Facilitate the Intended Uses?
Leanne Ketterlin Geller, Southern Methodist University; Lindsey Perry, Southern Methodist University; Katie Hogan, University of Texas, Austin

This paper presents findings from a qualitative study designed to examine score reports from universal screeners. We conducted a document analysis to identify common features, determine the degree of alignment with evidence-based report features, and how these features support the intended interpretations and uses of data from universal screeners.

Incorporating collateral information for reporting scores of social-emotional learning measures
Yang Wang, Education Analytics; Robert Meyer, Education Analytics; Andrew Rice, Education Analytics

This study applies several subscore augmentation methods to social-emotional learning instruments of California's CORE districts. Results show that the reliability of SEL scores can be increased via subscore augmentation techniques, especially at lower grades. Overall, augmented UIRT scores is the preferred method after considering CORE contexts.

Monday, April 16, 2018
8:15-10:15am, Manhattan, Coordinated Sessions, I8

Measurement Challenges in On-going Testing Environment: Potential Solutions

Session Chair: Suleyman Olgar, Florida Department of Education
Session Discussant: Ahmet Turhan, American Institutes for Research

This session brings four papers that address critical measurement challenges and potential solutions in ongoing educator certification testing. Each paper summarizes research findings and possible solutions in areas of differential item functioning, test speededness, reliability, and testing time limits. The first paper discusses challenges affecting reliability estimates and shares results from measurement strategies employed to address these challenges. The second paper summarizes a study to improve the performance of differential item functioning with data collection designs prevalent in professional certification testing programs. The third paper discusses implications of examination time limits in high-stakes testing as well as operational metrics by which to monitor for and determine test speededness. The fourth paper introduces a modified standard setting methodology to establish or validate computer based testing time limits in continues testing environment. This session presentations will provide potential solutions to major issues with on-going testing and help practitioners in their operational work to tackle these and similar issues.

Setting/Validating CBT Time Limits for Continuous Test Administrations: Modified Standard Setting Approach
Suleyman Olgar, Florida Department of Education

Reliability Estimates and Continuous Testing
Leah Kaira, ES Pearson

Small Sample Size DIF Challenges in on-going Testing
Alvaro Arce-Ferrer, Pearson

Test Speededness: Potential Solutions
Lauren White, Florida Department of Education

Monday, April 16, 2018
8:15-10:15am, Ambassador II, Individual Presentations, I9

Approaches to Assembly and Administration of Adaptive Tests

Session Discussant: Xiao Luo, Measured Progress

The Asymptotic Distribution of Mean Test Overlap Rate in Computerized Adaptive Testing
Edison Choe, Graduate Management Admission Council; Hua-Hua Chang, University of Illinois at Urbana-Champaign

Average test overlap rate is a standard measure of how secure a computerized adaptive test (CAT) is against compromise. A proof of its asymptotic distribution under random item selection is presented, thereby providing a theoretical baseline for assessing the potential security risk of a CAT design.

Using Position and Response Latency Constraints under Shadow Test Approach
Unhee Ju, Michigan State University; Shalini Kapoor, ACT, Inc; Yi-Fang Wu, ACT, Inc; Tony Thompson, ACT, Inc

Computerized adaptive testing may raise concerns about position effects and differential speededness, especially if using initial item parameter estimates from the paper-and-pencil testing. A simulation study was conducted to investigate these concerns by putting position and latency constraints in item selection under the shadow test approach.

Optimizing the Predicted Standard Error Reduction Stopping Rule in Computer Adaptive Testing
Scott Morris, Illinois Institute of Technology; Michael Bass, Northwestern University Feinberg School of Medicine; Elizabeth Howard, Illinois Institute of Technology; Richard Neapolitan, Northwestern University Feinberg School of Medicine

A computer adaptive test using the predicted standard error reduction (PSER) stopping rule will avoid unnecessary items by ending the exam if no items are likely to improve precision. This paper explores how to optimize the parameters of the PSER algorithm to obtain a desired tradeoff between precision and efficiency.

Strategies for Reducing Computation Time in Constrained Adaptive Testing with Shadow Tests
Alex Brodersen, University of Notre Dame; Wei He, NWEA

Shadow tests remain the popular choice for implementing content and other constraints within computerized adaptive testing, which require computation times that may be infeasible for large-scale online testing programs. This study presents a new method for reducing computation times and compares them to suggestions currently in the literature.

Monday, April 16, 2018
10:35am – 12:05pm, Majestic II, J1

We Can Do This: Communicating Information from Educational Assessments

Session Moderator: April Zenisky, University of Massachusetts Amherst
Session Moderator: Charles DePascale, National Center for the Improvement of Educational Assessment

The theme of this year's annual meeting is "Here and There and Back Again: Making Assessment a Stronger Force for Positive Impact on Teaching and Learning", and it is perhaps in the area of communicating test results that assessment has the most potential to have a direct and positive effect on teaching and learning. One can argue that reporting is the most important part of the assessment process. And yet, while there has been a marked surge in research on reporting in recent years, this aspect of testing remains a challenge for many agencies in several respects.

Reporting is a complex and multifaceted component of the testing process, with many opportunities for pitfalls as well as promise:

- what to report, how to report it and to whom to report what,
- understanding and embracing emerging mechanisms for dissemination of information from assessments to individuals and groups,
- issues in providing actionable, meaningful, and generalizable feedback, and
- supporting the appropriate use of information from assessments in educational practice.

In this session, the three speakers will provide their perspective on the past, present and future of reporting test scores/data and communicating information from assessments, with a focus on bridging research and practice to ensure that those efforts are aligned with and responsive to the various needs of intended users.

Supporting Effective Communication and Appropriate Use of Assessment Results
Diego Zapata-Rivera, Educational Testing Service

What's my Status? What's my Goal? What's my Pathway to that Goal?
Steven Ferrara, Measured Progress

Civis, smartphones and score reports
Howard Wainer, Princeton, NJ

Monday, April 16, 2018
10:35am-12:05pm, Broadway I, Coordinated Sessions, J2

New Insights on Survey Questionnaire Context Effects from Multiple Large-Scale Assessments

Session Chair: Jonas Bertling, Educational Testing Service
Session Discussant: Enis Dogan, National Center for Education Statistics

While survey questionnaires (SQs) are an established component of national and international large-scale assessments (LSAs) to assess noncognitive variables, a long history of research suggests that self-report responses may be influenced by a wide range of construct-irrelevant factors, including question format and question position. To ensure that conclusions based on SQ data are valid, it is important to better understand the mechanisms of context effects and their magnitude.

This symposium brings together five papers summarizing findings from several large pilots from the National Assessment of Educational Progress (NAEP) and the Programme for International Student Assessment (PISA) that systematically investigated context effects through carefully designed item format and sequence manipulations. Papers will focus both on context effects in terms of how items are arranged within booklets and with respect to semantic cues in the items. Results from these studies will help inform current and future questionnaire design and analysis practices, in terms of both practical and theoretical considerations. A discussant comment on each paper's relevance and implications for assembly of future SQs in LSAs.

Questionnaire Designs to Minimize the Impact of Context on Self-report
Paul Jewsbury, Educational Testing Service; Jonas Bertling, Educational Testing Service

Evaluating the Robustness of Survey Question Order Effects from Three Nationally-Representative Samples
Jan Alegre, Educational Testing Service; Paul Jewsbury, Educational Testing Service; Farah Qureshi, Educational Testing Service

Effects of Item Format (Discrete versus Matrix) on Grade 4 Student Responses
Debby Almonte, Educational Testing Service; Jonas Bertling, Educational Testing Service

Effect of Contextual Cue Placement on Survey Responses, Response Time, and Scalability
Farah Qureshi, Educational Testing Service; Jan Alegre, Educational Testing Service; Jonas Bertling, Educational Testing Service

Effects of Item Order and Gender on Anchoring Vignettes in 33 countries
Tamara Marksteiner, University of Mannheim; Eckhard Klieme, DIPF; Susanne Kuger, DIPF; Jonas Bertling, Educational Testing Service

Monday, April 16, 2018
10:35am-12:05pm, Broadway II, Coordinated Sessions, J3

Measuring instruction using classroom artifacts and portfolios: Evidence from four recent studies

Session Chair: Jose Felipe Martinez, University of California, Los Angeles
Session Discussant: James Pellegrino, University of Illinois at Chicago
Session Discussant: Courtney Bell, Educational Testing Service

High quality data about instructional practice in classrooms is essential for understanding and improving teaching and learning. Classroom artifacts and instructional portfolios are used widely to assess teaching for induction and certification (e.g., NBPTS, EdTPA), and there is growing interest in leveraging them to monitor instruction at scale. In principle these tools can offer rich, contextualized evidence for assessing teaching in greater depth than surveys, with lower cost and better coverage than classroom observations. However, little research systematically examines the technical properties, features, advantages and limitations of artifact-based measures of instruction. This session brings together cutting edge work from four teams using artifacts and portfolios to systematically assess features of classroom practice in different subjects and contexts. The applications represented in this session involve measures of instructional practice, content knowledge for teaching, and cognitive demand, in language arts, mathematics, and science classrooms, carefully examining the validity of inferences about instruction derived from evidence in artifacts and portfolios. The discussants will help outline a research agenda for advancing our knowledge about measures of instruction based on artifacts and portfolios, including approaches and sources of evidence for assessing the robustness of the instruments, and the validity of the measures.

Examining the Validity and Impact of Beginning Teacher Assessment & Accountability Systems
Raymond Pecheone, Stanford University; Kevin Bastian, University of North Carolina

Measuring science instruction using a tablet portfolio app: Reliability and Validity
Jose Felipe Martinez, University of California, Los Angeles; Jayashri Srinivasan, University of California, Los Angeles; Brian Stecher, The RAND Corporation; Amanda Edelman, The RAND Corporation; Matt Kloser, The University of Notre Dame; Matt Wilsey, The University of Notre Dame

Measuring Content Knowledge for Teaching using Instructional Artifacts
Robert Zisk, Rutgers University; Drew Gitomer, Rutgers University; Eugenia Etkina, Rutgers University

Students thinking writing: Validity of inferences of enactment of cognitively demanding tasks
Richard Correnti, University of Pittsburgh; Lindsay Clare-Matsumura, University of Pittsburgh

Monday, April 16, 2018
10:35am-12:05pm, Broadway III, Coordinated Sessions, J4

Innovative Approaches to Standard Setting: Responding to a Changing Assessment Environment

Session Chair: Douglas Becker, Houghton Mifflin Harcourt
Session Discussant: Gregory Cizek, University of North Carolina-Chapel Hill

The enactment of the Every Student Succeeds Act (ESSA) is certainly having an impact on how states deal with assessment and accountability. ESSA, however, did not change the mandate that states employ accountability systems to monitor school performance. States have flexibility in the format of assessment; assessments may include portfolios, projects, or extended performance tasks; states can use computer-adaptive assessments; states must provide understandable information to parents. As assessments and assessment systems evolve and adapt under ESSA, so too must the methodologies associated with determining cut scores and proficiency levels. In this session, the presenters will describe and discuss three approaches to standard setting that help practitioners respond to a changing assessment environment. The discussant will draw on his extensive experience to offer comments and insights on the approach, methodology, and possible policy implications associated with each of the presentations. The alignment of new and innovative approaches to standard setting that respond to the added flexibility introduced with ESSA around assessments and accountability hold the promise of enhancing the validity of large-scale assessment programs while addressing issues of assessment mode and/or format at the root.

The Essay Profile Method: An Innovative Approach in Standard Setting
Liru Zhang, Delaware Department of Education

Interval Validation Method: Achievement Level Setting Based on Large Pools of Items
William Insko, Jr., Houghton Mifflin Harcourt

Efficacy of Engineered Cut Scores: Embedding Standard Setting in Principled Assessment Design
Daniel Lewis, ACT, Inc.

Monday, April 16, 2018
10:35am - 12:05pm, Ambassador III, Individual Presentations, J5

Emerging Research on the Adaptation of Adaptive Tests

Session Discussant: Tim Davey, Educational Testing Service

Evaluating Indicators of the Amount of Adaptation to 3PL Computerized Adaptive Test
Sewon Kim, Michigan State University; Unhee Ju, Michigan State University; Mark Reckase, Michigan State University

This study investigates indicators of the amount of adaptation to 3PL CATs. Simulation studies were conducted depending on item pool size, variation of the item parameters in the pool, and exposure control procedure. Based on the results, guidelines for interpreting the indicators are discussed.

Multistage Testing Routing Designs: Adjacent vs. Nonadjacent
Han Yi Kim, Measured Progress; Hyung Jin Kim, The University of Iowa

For multistage testing, examinees sometimes can only proceed to modules in the next stage that are adjacent to the current module. However, from fairness perspectives, examinees should also be given chances to move to nonadjacent modules. This study investigates the impact of routing designs on performance classification accuracy.

What Information Works Best? A Comparison of Routing Methods
Halil Ibrahim Sari, Kilis 7 Aralik University; Anthony Raborn, University of Florida

There are many item selection methods proposed for computerized adaptive testing applications. However, not all of them have been used in computerized multistage testing (CMT). The main purpose of this study is to examine performance of these methods when they are used as a routing method in CMT framework.

Discontinuation Rules in Testing: New Results on Ignorability, Local Dependency, and Bias
Matthias von Davier, NBME; Youngmi Cho, Pearson; Tianshu Pan, Pearson

New results on discontinue rule scoring used in intelligence testing are presented. The Stanford-Binet test and Kaufman Assessment Battery are examples using this rule. We present new results on ignorability of missingness under discontinue scoring rule, and show that operational scoring rule introduce bias and local dependency.

Setting a Time Limit for a Computer Adaptive Test
Furong Gao, Pacific Metrics Corporation; Kyoko Ito, Defense Personnel Assessment Center; Daniel Segall, Defense Personnel Assessment Center

As the current computer-adaptive test (CAT) of English proficiency undergoes changes, the present testing time needed to be re-evaluated. Using item response latency data from the current CAT test, this real-data simulation study estimated the testing time required for the new CAT test. R-code will be made available.

Monday, April 16, 2018
10:35am - 12:05pm, Majestic I, Individual Presentations, J6

Advances in Estimation of DCM

Session Discussant: Laine Bradshaw, UGA

Insights from reparameterized CDMs: Implementation, Monotonicity, and Duality

Lawrence DeCarlo, Teachers College, Columbia University

Reparameterizations of CDMs provide a simple way to fit the models with standard software. They also provide insights into concepts such as monotonicity and how to implement it in practice, and clarify aspects about the duality of the DINO and DINA models, resulting in a deeper understanding of these models.

A Multilevel Logistic-Hidden Markov Model for Learning under Cognitive Diagnosis

Susu Zhang, University of Illinois at Urbana-Champaign; Hua-Hua Chang, University of Illinois at Urbana-Champaign

We propose a multilevel logistic hidden Markov model for learning, with the learner's previous mastery of other skills and the effectiveness of the learning material as covariates of learning outcome. An MCMC algorithm is propocogsed for parameter estimation, and a simulation study is conducted to evaluate model parameter recovery.

Loglinear cognitive diagnosis model estimation via particle swarm optimization

Zhehan Jiang, University of Kansas

A loglinear cognitive diagnosis model (LCDM) is estimated via a global optimization approach- particle swarm optimization (PSO), which is an efficient method of handling local maxima problem. The application of the PSO to LCDM estimation is introduced, explicated, and evaluated via a Monte Carlo simulation study in this paper.

The Performance of the Constraint-weighted Item Selection Procedures in Variable-length CD-CAT

Ya-Hui Su, National Chung Cheng University; Hua-Hua Chang, University of Illinois at Urbana-Champaign

There are only a limited numbers of previous studies examining how to optimally construct cognitive diagnostic computerized adaptive testing (CD-CAT). It is challenging to meet various constraints for test construction. This study investigated the constraint-weighted version of posterior-weighted cognitive discrimination index and posterior-weighted attribute-level discrimination index in variable-length CD-CATs.

Cognitive Diagnostic Modeling of Responses and Time for Items in Multiple Contexts

Hong Jiao, University of Maryland, College Park; Peida Zhan, Beijing Normal University; Dandan Liao, University of Maryland, College Park; Kaiwen Man, University of Maryland, College Park

This study proposes a joint modeling approach of item responses and response time for innovative items embedded in multiple context for cognitive diagnosis. Model parameter estimation is explored. Model parameter recovery is evaluated under different simulated study conditions. The effect of ignoring complex local item dependence is examined.

Monday, April 16, 2018
10:35am - 12:05pm, Plymouth, Individual Presentations, J7

Exploring Speededness: Detection and Impact

Session Discussant: Brian Clauser, National Board of Medical Examiners

Using CUSUM and CPA Method to Detect Speededness

Xiaofeng Yu, University of Notre Dame; Ying Cheng, University of Notre Dame

In this study, we conduct a comprehensive comparison of twelve cumulative sum and three change-point analysis procedures in detecting test speededness. Two speededness mechanisms are considered to test robustness and flexibility of the two methods. Simulation studies show that L, S and T7 performed better based on a summative comparison.

Exploring Characteristics of Speeded Examinees Using LDA

Meereem Kim, University of Georgia; Allan Cohen, University of Georgia

Covariates can be used to help explain characteristics of latent classes obtained from a mixture IRT model. In this study, LDA is used to construct covariates from constructed response answers for latent class membership between speeded and nonspeeded examinees detected by a mixture IRT model.

Using Rapid Responses to Evaluate Test Speededness

Richard Feinberg, National Board of Medical Examiners; Daniel Jurich, National Board of Medical Examiners

Several recent studies have proposed novel methods for determining rapid responses by utilizing the conditional probability of correct response given response time. The current study furthers this research by comparing different rapid response definitions and how identifying rapid responses can help assess the impact of speededness.

Using timing data to evaluate differential warm up effects in NAEP

Ruhan Circi, American Institutes for Research; Young Yee Kim, American Institutes for Research

The introduction of digitally-based assessments (DBA) raises concerns about the potential impact of differential warm-up effects on performance. Warm-up effects exist if students tend to spend more time than they actually need, at the beginning. This study aims to investigate the relationship between warm-up effects and performance in NAEP DBA.

Exploring Issues in Test Speedness: Insight Gained from an Early Literacy Assessment

Qinjun Wang, University of Minnesota - Twin Cities; Michael Rodriguez, University of Minnesota - Twin Cities; Kristin Running, University of Minnesota - Twin Cities; Alisha Hollman, University of Minnesota - Twin Cities; Scott McConnell, University of Minnesota - Twin Cities

This study investigated the stability and functionality of response speed parameter in an untimed literacy assessment. Under the Rasch modeling methods, a fundamental response speedness assumption was examined across five early literacy domains and a potential test taker response speed parametrization was investigated.

Monday, April 16, 2018
10:35am - 12:05pm, Manhattan, Individual Presentations, J8

Scoring with Multiple Categories

Session Discussant: Jiyoon Park, Federation of State Boards of Physical Therapy

Awarding Partial Credit: A New Approach to Student Learning Objectives (SLO)

Pamela Wong, New York City Department of Education; Joseph Jensen, New York City Department of Education; Jordan Munn, New York City Department of Education

A large public school district is exploring how to improve its Student Learning Objectives (SLO) methodology by experimenting with awarding partial credit in order to make it fairer and more consistent across teachers and schools. The paper will meticulously describe the methodology and report simulated results based on 2016-2017 data.

Fair partial-credit scoring of sentence-sequencing tasks to assess second language reading

Karen Dunn, British Council; Richard Spiby, British Council

This paper addresses fair score allocation for sentence-sequencing items in L2 reading tests. Operational testing data were re-scored using different partial-credit approaches. Results from IRT, CFA and correlation analyses are presented with recommendations for a solution that reflects the underlying construct well, and makes best use of interim score categories.

The Delta-Scoring Method Adapted for Polytomous Test Items

Dimiter Dimitrov, National Center for Assessment in Saudi Arabia; Yong Luo, National Center for Assessment in Saudi Arabia

An approach to scoring tests with binary items, called *delta scoring*, is used at the National Center for Assessment in Saudi Arabia. As some tests include polytomous items, this study provides an approach to delta-scoring of such items and parallels the results with those obtained under the Graded Response Model.

An Item Response Tree Model for Validating Rubric Scoring Processes

Aaron Myers, James Madison University; Allison Ames, James Madison University

Performance assessments are often considered a more authentic alternative to measuring higher-order skills. However, inconsistent application of rubrics may cause raters to differ on scores ascribed to a given performance. IRTree models were utilized to validate raters' scoring processes when using a traditional rubric and an alternative logic-based scoring method.

Monday, April 16, 2018
10:35am - 12:05pm, Ambassador II, Individual Presentations, J9

Test Score Use, Stakeholder Perceptions, and Evidence of Consequences

Session Discussant: Stephen G. Sireci, University of Massachusetts Amherst

In the eye of the beholder: Stakeholder perceptions of validity evidence
Michelle Croft, ACT, Inc.; Paul Nichols, ACT, Inc.; Emily Lai, Pearson

The study examines differences in the types of validity evidence the five stakeholder groups (parents, teachers, policymakers, lawyers, and psychometricians) find relevant and how the responses change based on score use (low stakes, high stakes for students, and high stakes for teachers).

A Framework to Support Validation and Evaluation: Performance assessment applications
Brian Gong, Center for Assessment; Yuxi Qiu, University of Florida (Gainesville)

Performance assessments offer a direct way for students to demonstrate what they can do, often in "real-world" settings. However, developing and managing validation evidence is challenging. To surmount this challenge, this study suggests a framework that emphasizes close attention to closely defining user intentions in validation, and evaluation of consequences

Breaking the rules: Validation when the purpose changes but the test doesn't
Susan Davis-Becker, ACS Ventures, LLC; Ellen Forte, edCount LLC

At present, there are numerous examples of using existing tests for new purposes such as a college readiness assessment being used as a statewide high school assessment. This paper considers the validity evidence necessary to support such applications of measures for new purposes.

Validity issues in large scale test use contexts: Analyzing stakeholder perspectives
Madhabi Chatterji, Columbia University, Teachers College; Meiko Lin, Columbia University, Teachers College

Few studies portray multiple perspectives on recurring validity issues that arise in large scale testing contexts. This qualitative research study addresses this gap in the measurement and policy literature by systematically analyzing stakeholder blogs to catalog the specific types, causes, and consequences of "validity" challenges in visible assessment programs.

Consequential Validation: Where are we after 25 Years of Effort?
John Poggio, University of Kansas; Susan Lyons, NCIEA; Peter Ramler, University of Kansas

The paper reviews and evaluates published studies and research of the past 25 years as to the value and viability of Consequential Related Validity evidence in education and psychology. A meta-review of consequential validation work (75+ to date) is provided. We offer guidance as to future needs and direction.

Monday, April 16, 2018
10:35am-12:05pm, Gershwin I, Graduate Student Research Session, J10

GSIC Graduate Student Poster Session 2

Electronic Board #1

Effects of Item Positions and Psychological Factors on Item Parameter Estimation

Nayeon Yoo, Teachers College, Columbia University; Ummugul Bezirhan, Teachers College, Columbia University; Young-Sun Lee, Teachers College, Columbia University

The purpose of the study was to examine the effects of item positions and psychological factors on item parameter estimation via item response theory (IRT) and structural equation modeling (SEM). Real-world data analyses were conducted using TIMSS 2015 data, and simulation studies were conducted to examine the recovery of parameters.

Electronic Board #2

Evaluation of Validity Claims for the Perceived Stress Scale

David Alpizar, Washington State University; Thao Vo, Washington State University; Brian French, Washington State University; Scott Plunkett, California State University Northridge

Validity claims for the Perceived Stress Scale (PSS) were evaluated. Factor structure claims between ethnic and gender groups were examined. Associations with external variables were evaluated for PSS scores with quality of life, health, anxiety, and depression outcomes for university students. Factorial invariance and expected relationships between variables were supported.

Electronic Board #3

Calibration of Automatic Generated Items Using an Item Modeling Approach

Yu Bai, Teachers College, Columbia University; Andrew Dallas, National Commission on Certification of Physician Assistants; Fen Fan, National Commission on Certification of Physician Assistants; Joshua Goodman, National Commission on Certification of Physician Assistants

This study examines how to safely and appropriately calibrate items from item families created by Automatic Item Generation (AIG) methods within a certification/licensure context. The study compares the accuracy of item parameter estimation under eight multi-level Bayesian IRT models in a simulation study.

Electronic Board #4

Handling Missing Data with Imputation in Cognitive Diagnostic Models

Ummugul Bezirhan, Teachers College, Columbia University; Yu Bai, Teachers College, Columbia University; Young-Sun Lee, Teachers College, Columbia University

Missing data may pose challenges in estimation accuracy, the generalizability of results and statistical power in educational and psychological measurement. This study examines the effects of traditional, single and multiple imputation techniques on parameter estimates and classification accuracy of DINA model under both ignorable and non-ignorable missing data.

Electronic Board #5

A Multigroup Testlet Model for Cognitive Diagnosis

Dandan Liao, University of Maryland; Hong Jiao, University of Maryland; Peida Zhan, University of Maryland

This study proposes a multigroup testlet diagnostic classification model within the logistic cognitive diagnostic model framework, which accommodates differences among student populations in testlet-based tests. The proposed model is evaluated through a simulation study with respect to group difference in attribute profiles, testlet variance, and differential item functioning (DIF) magnitude.

Electronic Board #6

Look-ahead Content Balancing Method in Variable Length Computerized Classification Testing

Xiao Li, University of Illinois at Urbana-Champaign; Hua-Hua Chang, University of Illinois at Urbana-Champaign

Look-ahead content balancing method (LA-CB) is proposed as a feasible solution to balancing content areas in variable length computerized classification testing (VL-CCT). Integrated with heuristic item selection methods, LA-CB method will be evaluated with respect to classification accuracy, content area balancing, and exposure control.

Electronic Board #7

Using a diagnostic model pretest to evaluate mathematics skills in middle school

Peter Ramler, The University of Kansas; David Livingston, Most Pure Heart of Mary Catholic School; Jonathan Templin, The University of Kansas

Teachers are often frustrated in finding a useful assessment to diagnose the skill level of students at the start of a teaching unit. What is needed is an assessment that is easy to administer and produces straight forward results. Diagnostic Classification Models are capable of fulfilling these needs.

Electronic Board #8

An application of automatic item generation on a course exam using R

Yating Zheng, University of Maryland, College Park

This study explores the application of automatic item generation (AIG) on a course exam. 1-layer item model is used to generate items. Rasch model is used to estimate the psychometric properties of the generated items. AIG enables rapid and efficient generation of large numbers of items and improves test security.

Electronic Board #9

Automated scoring and feedback system in computer-based literacy assessment: Graphic organizer creation

Hyunah Kim, University of Toronto; Clarissa Lau, University of Toronto; Megan Vincett, University of Toronto; Eunice Jang, University of Toronto

As part of a broader project developing a formative computer-based literacy assessment for elementary students in Ontario, this study focuses on building and validating an automated scoring and feedback system for graphic organizer creation tasks. The system attempts to assess both cognitive and non-traditional intrapersonal constructs that relate to literacy.

Electronic Board #10

The Impact of Subpopulation Item Parameter Drift on Equating

Liuhan Cai, University of Nebraska-Lincoln

Subpopulation item parameter drift (SIPD) concerns the change of item parameters over time that is only specific to subpopulations. This study examines the impact of SIPD on anchor test dimensionality and scaling coefficients under conditions of ability difference, number of anchor items displaying drift, and magnitude of drift.

Electronic Board #11

SAT Reading Construct Validation: Predicting Item Difficulty from Text and Item Complexity

Maryam Pezeshki, Georgia Institute of Technology; Clifford Hauenstein, Georgia Institute of Technology; Susan Embretson, Georgia Institute of Technology

The Linear Logistic Model (LLTM, Fischer, 1973) was applied to SAT reading test items to model difficulty from both text complexity and item complexity indices. Results showed significant predictions by the cognitive complexity variables. Implications for item design and item banking by complexity sources are discussed.

Electronic Board #12

The Cubic B-Spline Presmoothing Method under the CINEG Design

Widad Abdalla, University of Iowa

In Equating, smoothing is designed to reduce random error without introducing too much systematic error. The purpose of this study is to introduce and compare the Cubic B-Spline Method to the Cubic Spline Postsmoothing Method, LogLinear Presmoothing Method, and no smoothing under the common item non-equivalent groups (CINEG) design.

Electronic Board #13

Understanding PISA Collaborative Problem-Solving Assessment: A Cross-country Comparison

Shuang Wang, University of Wisconsin - Milwaukee; Bo Zhang, University of Wisconsin - Milwaukee

PISA administered collaborative problem-solving assessment for the first time in 2015. We study whether this measure of collaborative cognitive processes is invariant across countries. We also investigate its convergent validity by exploring its relationship with other cognitive assessments.

Electronic Board #14

Exploring Metropolis-Hasting Robbins-Monro Estimation Method in MIRT Models under Multiple-group Concurrent Calibration

Ye Ma, the University of Iowa; Won-Chan Lee, the University of Iowa; Stephen Dunbar, the University of Iowa

Current research aims to explore the MH-RM algorithm (Cai, 2010a) in the context of multiple-group concurrent calibration and to compare its performance with the traditional EM method using various multidimensional IRT models in a simulation study.

Electronic Board #15

Measurement Equivalence of a Student Experiences Survey Across Contrasting Pairs of Universities

Daniela Cardoza, University of Iowa; Thapelo Ncube, University of Iowa; Robert Ankenmann, University of Iowa

Measurement equivalence/invariance (ME/I) has been studied using confirmatory factor analysis (CFA) and item response theory (IRT). The purpose of this study is to apply CFA and IRT to the assessment of ME/I in the 2016 Student Experiences in the Research University Survey across contrasting pairs of schools.

Electronic Board #16

A Computerized Adaptive Testing Exposure Method for Cognitively-Based Multiple- Choice Assessment

Hulya Duygu Yigit, University of Illinois at Urbana-Champaign; Miguel Sorrel, Universidad Autónoma de Madrid; Juan Barrada, Universidad de Zaragoza; Jimmy de la Torre, The University of Hong Kong

Jensen-Shannon divergence item selection index produced promising results with polytomous data, but an uneven usage of the item pool in CD-CAT. In this paper, different item exposure methods including the modified progressive, modified proportional, and item-eligibility methods are compared under the *multiple-choice deterministic inputs, noisy "and" gate* model setting.

Electronic Board #17

Introducing a New Item Fit Index for Multiple-Choice DCMs

Yanyan Fu, The University of North Carolina at Greensboro; Robert Henson, The University of North Carolina at Greensboro

An item fit index(Q-index) for multiple-choice DCMs was proposed and studied using simulation methods. Various of conditions were manipulated including sample size, type of misspecified Q-matrix, and proportion of items that have misspecified Q-matrix. The results show that Q-index out-performed an existing fit index proposed by DiBello et al. (2015).

Electronic Board #18

Posterior Predictive Model Checking of Local Misfit for Bayesian Confirmatory Factor Analysis

Chi Hang Au, James Madison University; Allison Ames, James Madison University

Posterior predictive model checks (PPMC) are one Bayesian approach to model-data fit. PPMC global fit has been the focus of Bayesian CFA applications, ignoring the nuanced information in local misfit diagnostics. This study develops a PPMC approach for local misfit and applies it to a categorical scale on motivation.

Electronic Board #19

Using Public Data to Examine Potential Effects of an ACT-for-All Policy

Francis O'Donnell, University of Massachusetts, Amherst

The purpose of this study was to investigate changes in postsecondary enrollment in Louisiana following the implementation of a policy that required all public high school juniors to take the ACT. Data from three pre-policy and two post-policy student cohorts were analyzed. Methodological considerations for future research are discussed.

Electronic Board #20

Measurement Invariance and Predictive Validity of a School Culture and Climate measure

Leon Gilman, University of Wisconsin - Milwaukee

School culture and climate measures are generally designed for older students. This research investigates whether a popular school culture and climate measure holds measurement invariance between elementary and middle school students. Results show perceptions of the learning environment are multidimensional, measurement invariance does not hold, and predicts academic achievement.

Electronic Board #21

Exploring the Rating of Cognitive Complexity in Mathematics Assessment Items

Deborah La Torre, UCLA

This research is designed to lay the groundwork for development of a new or revised framework to analyze the cognitive complexity of mathematics assessment items. This will be accomplished using a mixed methods approach including literature reviews, descriptive analyses of existing complexity ratings, and cognitive labs with students and teachers.

Electronic Board #22

Considering sampling errors in estimating Value-Added Ratios of subscores: A bootstrap method.

Duy Pham, University of Massachusetts Amherst

This study develops an existing method to justify added value of subscores by using bootstrap to consider sampling errors of Value-Added Ratios of individual and institutional sub-scores. Preliminary findings suggested that the bootstrap was implementable and the VARs seemed to be very precise. Future directions are discussed.

Electronic Board #23

Supervised Text Analysis for Mixture Groups

Seohyun Kim, The University of Georgia; Zhenqiu Lu, The University of Georgia; Allan Cohen, The University of Georgia

Supervised Latent Dirichlet Allocation (Supervised LDA) can be used to jointly model text data and related labels such as scores of answers to constructed response (CR) items. We extend the supervised LDA to find topics underlying responses to CR items and investigate the relationship between topic proportions and scores.

Electronic Board #24

Assessing the Dimensionality Assumption under Data Generating and Analysis Model Mismatch

Kirsten Hochstedt, Penn State University

The performance of select IRT dimensionality assessment methods when guessing behavior was included in the generated responses, but not the analysis model, were compared. A simpler analysis model should reduce the number of examinees and items required for accurate parameter estimation. Four NOHARM-based methods that detect dimensionality violations were compared.

Electronic Board #25

Measurement Invariance in Noncognitive Measures: Validity Approach Using Explanatory Item Response Modeling

Jose Palma, University of Minnesota; Youngsoon Kang, University of Minnesota; Okan Bulut, University of Alberta; Michael Rodriguez, University of Minnesota

Using a partial-credit explanatory item response model, we examine validation concerns of score interpretation of two developmental measures, Social Competence and Empowerment across grade levels. We find significant interactions between items and grade levels for both constructs implying different perceptions of items across grades and potentially influencing score interpretation.

Monday, April 16, 2018
12:25 – 1:55, Majestic 2, Invited Session, K1

Measurement Problems – A look back to help us look ahead
Measurement Problems Session 3
Session Moderator: Phoebe Winter, consultant
Panelist: Derek Briggs, University of Colorado Boulder
Panelist: Andrew Butler, Washington University in St. Louis
Panelist: Ellen Forte, edCount
Panelist: Kathleen Scalise, University of Oregon
Panelist: Sandip Sinharay, Educational Testing Service

The problems mentioned are merely samples of problems, yet they will suffice to show how rich, how manifold and how extensive the ... science of [measurement] today is, and the question is urged upon us whether it is doomed to the fate of those other sciences that have split into separate branches, whose representatives scarcely understand one another, and whose connection becomes ever more loose. I do not believe this nor wish it. [Measurement] is in my opinion an indivisible whole, an organism whose vitality is conditioned upon the connection of its parts. For with all the variety of . . . knowledge, we are still clearly conscious of the similarity of the logical devices, the relationship of the ideas in [measurement] as a whole and the numerous analogies in its different departments.

(Wainer's 1993 conclusion, adapting Hilbert, 1902, p. 477)

In this third Measurement Problems session, panelists will examine unsettled questions that vex us today and identify new measurement challenges that are emerging or are likely to emerge in a world in which the walls between assessment and instruction/learning are being broken down; a world of personalized instruction with a focus on the individuals' learning and growth.

Monday, April 16, 2018
12:25-1:55pm, Broadway I, Coordinated Sessions, K2

Validity and Diversity Challenges in Post-Secondary Admissions

Session Chair: David Klieger, Educational Testing Service
Session Chair: Brent Bridgeman, Educational Testing Service
Session Discussant: Wayne Camara, ACT

This coordinated session addresses perennial challenges post-secondary admissions officers face in selecting students likely to achieve academic performance objectives while simultaneously admitting a diverse group of applicants. U.S. law increasingly has complicated use of affirmative action, such that many higher education institutions struggle to achieve their student diversity objectives as they look for information indicating which applicants are likely to be academically successful. Four presenters each will present a measurement-based approach to better understand and help address these challenges.

Holistic vs. Statistically-Based Decision-Making: Measuring Accuracy and Diversity
David Klieger, Educational Testing Service; Paola Heincke, Educational Testing Service; Travis Liebtag, Educational Testing Service; Steven Holtzman, Educational Testing Service; Nimmi Devasia, Educational Testing Service; Adam Bacall, Educational Testing Service

Non-cognitive assessments increase admissions diversity and quality: A pilot study
Patrick Kyllonen, Educational Testing Service

Increasing the College Admissions Rate for Students from Low-Income Neighborhoods
Rebecca Zwick, Educational Testing Service; Lei Ye, Educational Testing Service; Steven Isham, Educational Testing Service

Using Single- versus Multi-hurdle Approaches in Higher Education Admissions
Guangming Ling, Educational Testing Service; Jennifer Minsky, Educational Testing Service; Zhitong Yang, Educational Testing Service

Monday, April 16, 2018
12:25-1:55pm, Broadway II, Coordinated Sessions, K3

Score Reporting for High-Stakes Certification and Licensing Programs

Session Chair: Andrea Gotzmann, Medical Council of Canada
Session Discussant: Ronald Hambleton, University of Massachusetts

In recent years, score reporting research has had an increased presence at NCME, with focus on evaluating and improving score reports and subscore calculations. Most research has been in the context of K-12 where criterion- and norm-referenced score interpretations are supported. Less score reporting research has focused in the credentialing/licensing context, where the primary criterion-referenced score interpretation is a pass/fail status, yet examinees desire enhanced feedback for skill improvement and retaking the exam after a failure, and stakeholders want to evaluate outcomes for other purposes such as program evaluation.

The focus of this symposium is score reporting in a licensing/credentialing context. The four papers will cover: (1) eliciting feedback from examinees to improve score reports, (2) augmented subscores for institutional reporting, (3) evaluating format and understanding of score reports with examinees, and (4) eliciting feedback on two types of score reports (examinee level and aggregate level score reports) from examinees and other stakeholders. This research can be applied to other licensing/credentialing organization contexts. Most importantly, while the focus is on score reports for licensure/certification, many of the aspects discussed in this symposium are also relevant and applicable to score reporting in K12 and other high education contexts.

Score Reports: a Collaborative Design between Measurement, Communications, and Subject Matter Experts-Test
Bradley Brossman, American Board of Internal Medicine

Exploring the Score Report of a Computerized Adaptive Testing Program
Ada Woo, NCSBN; Wei Xu, NCSBN; Hong Qian, NCSBN

Providing Actionable Feedback on a High Stakes Licensure Examination
Amanda Clauser, National Board of Medical Examiners

A Model-Based Plan for Evaluating a High-Stakes Medical Licensure Exam's Score Reports
Ramsey Cardwell, University of North Carolina Greensboro; Andrea Gotzmann, Medical Council of Canada; Cecilia Alves, Medical Council of Canada; Liane Patsula, Medical Council of Canada; André De Champlain, Medical Council of Canada

Monday, April 16, 2018
12:25-1:55pm, Broadway III, Coordinated Sessions, K4

Peeking into Student Writing Behaviors in NAEP: Why and How

Session Chair: Yue Jia, Educational Testing Service
Session Chair: Mo Zhang, Educational Testing Service
Session Discussant: Eunice Greer, National Center for Education Statistics

Recently, the NAEP writing assessments begin to collect writing process data, including keystroke logs. This allows for broader and more in-depth descriptions of student writing performance. In this symposium, we will present our latest research and development on the use of writing process data in NAEP writing.

Value and Considerations for Collecting Writing Process Logs in Writing Assessments
Paul Deane, Educational Testing Service; Mo Zhang, Educational Testing Service; Gary Feng, Educational Testing Service; Hillary Persky, Educational Testing Service; Robert Finnegan, Educational Testing Service; Patricia Donahue, Educational Testing Service; Jie Gao, Educational Testing Service

Analyzing Writing Process in NAEP Writing Assessment: Implementation and Evidence Extraction
Gary Feng, Educational Testing Service

Psychometric Considerations for Analyzing Process Data in NAEP Writing
Yi-Hsuan Lee, Educational Testing Service; Jie Gao, Educational Testing Service; Yue Jia, Educational Testing Service

Exploring Subgroup Differences Using Writing Process Indicators
Jie Gao, Educational Testing Service; Gary Feng, Educational Testing Service

Monday, April 16, 2018
12:25-1:55pm, Ambassador III, Individual Presentations, K5

Issues in Growth Modeling

Session Discussant: Anton Beguin, CITO

Evaluating multiple imputation for estimating nationally representative student growth percentiles
Jeff Allen, ACT, Inc.

This study evaluates an imputation-based method for estimating nationally-representative student growth percentiles (SGPs). School and student variables observed both for the sample and population are the basis for imputation. Using simulation, the performance of imputation-based and standard methods are evaluated under various missing data scenarios.

The Sensitivity of Test-Based School Growth Metrics to Transformations of Scale
Darrick Yee, Harvard Graduate School of Education; Andrew Ho, Harvard Graduate School of Education

This paper presents a framework for modeling variation in key parameters of many growth-based accountability models in a unified manner. We employ this to examine the sensitivity of growth rankings to changes in model features, which would otherwise be impractical due to the diversity of available models and assessments.

Constraints arising from a decelerating logarithmic growth in mathematics proficiency
David Andrich, The University of Western Australia; Ida Marais, The University of Western Australia

This paper characterises the decelerating growth in proficiency in mathematics on a measurement scale as a near-perfect logarithmic function of years in schooling. Implications of this function summarise *efficiently and powerfully* a range of studies concerned with beginning a growth trajectory of mathematical proficiency in early childhood.

Accounting for Test Score Measurement Errors in Student Growth Models
Pei-Hsuan Chiu, University of Connecticut; H. Jane Rogers, University of Connecticut; Hariharan Swaminathan, University of Connecticut

This study compares three approaches for modeling student growth that take into account heteroscedastic test score measurement errors. Accuracy and standard errors of prediction among the three models were compared.

Predictive Validity of Classroom Observations and Self-Reflections for Student Growth
Linda Reddy, Rutgers, the State University of New Jersey; Adam Lekwa, Rutgers, the State University of New Jersey; Kevin Crouse, Rutgers, the State University of New Jersey; Christopher Dudek, Rutgers, the State University of New Jersey; Ryan Kettler, Rutgers, the State University of New Jersey; Ilona Arnold-Berkovit, Rutgers, the State University of New Jersey; Jlefang Hu, Rutgers, the State University of New Jersey; Anh Hua, Rutgers, the State University of New Jersey; Alex Kurz, Arizona State University

This investigation examined the predictive validity of the Classroom Strategies Assessment System, a multi-rater assessment of teacher practices, to growth scores on the PARCC assessment. Results indicated principal observation and teacher self-report predict growth in achievement. Implications for future practice and research will be discussed.

Monday, April 16, 2018
12:25-1:55pm, Majestic I, Individual Presentations, K6

Proficiency Estimation

Session Discussant: Hong Jiao, University of Maryland College Park

Elimination versus correction for guessing in multiple-choice exams: an empirical comparison
Rianne Janssen, KU Leuven; Qian Wu, KU Leuven; Jef Vanderoost, KU Leuven; Tinne De Laet, KU Leuven

Elimination scoring is an alternative way of administering multiple-choice exams that discourages guessing other than through the treat of receiving a penalty for wrong responses. It encourages students to express their partial knowledge. The difference in results with correction for guessing is small for students who master the topic well.

Tailored Booklets: Improved Estimates of Latent Traits in Large-Scale Assessment?
Leslie Rutkowski, University of Oslo; David Rutkowski, University of Oslo; Yuan-Ling Liaw, University of Oslo; Tyler Matta, Amplify Education and University of Oslo

The inclusion of easy booklets in international assessments is intended to strengthen proficiency estimation at the lower end of the spectrum. We examine whether the current PISA designs that include less difficult test items can improve proficiency estimation. Findings suggest that design-based solutions must change to accurately capture low proficiency.

Bayesian Extension of Biweight and Huber Weight for Robust Ability Estimation
Hotaka Maeda, University of Wisconsin-Milwaukee; Bo Zhang, University of Wisconsin-Milwaukee

Based on the popular Biweight and Huber weight, two new Bayesian robust ability estimation methods were developed. The prior distribution can compensate the information lost due to aberrant responses. It can also reduce the detrimental effects from downweighting the non-aberrant responses. Overall, the new methods improve ability estimation accuracy.

Improving Psychometric Precision through CFA Modeling of Transient Error and Scale Coarseness
Walter Vispoel, University of Iowa; Carrie Morris, University of Iowa; Murat Kilinc, University of Iowa

We demonstrate how CFA models can quantify effects of transient error and scale coarseness on reliability and disattenuated validity coefficients. Results showed that reliability was overestimated when ignoring transient error and underestimated when ignoring scale coarseness. These factors along with correlated transient errors within occasions strongly affected disattenuated validity coefficients.

Impact of Composite Population Priors on Proficiency Estimates from Computer Adaptive Tests
Kristin Morrison, ACT, Inc.; Susan Embretson, Georgia Institute of Technology

Bayesian priors can provide more examinee information in ability estimation, but may result in biased estimates if inappropriate. This study conducted a simulation to examine the implications of group-based priors on ability estimation in an educational context. Results suggest that disadvantageous consequences may be observed for various ability subgroups.

Monday, April 16, 2018
12:25-1:55pm, Plymouth, Individual Presentations, K7

Processes and Considerations in Adaptive Test Assembly

Session Discussant: Kimberly F. Colvin, SUNY Albany

Automatic Detection of Enemy Item Pairs Using Latent Semantic Analysis
Fang Peng, University of Illinois at Chicago; Xiao Luo, National Council of State Boards of Nursing; Hong Qian, National Council of State Boards of Nursing; Ada Woo, National Council of State Boards of Nursing

Latent Semantic Analysis (LSA) offers computational methods for extracting and representing the meaning of words as underlying dimensions of a large text corpus. This paper presents an automatic approach of using LSA to measure item similarity with the goal of identifying enemy relationship in item banks.

Enemy Item Detection with Natural Language Processing: Latent Dirichlet Allocation
J. B. Weir, University of North Carolina at Greensboro; Andrew Dallas, NCCPA; Joshua Goodman, NCCPA; Fen Fan, NCCPA

Enemy items, by virtue their similarity to one another, can distract examinees from intended constructs or clue them to correct answers to subsequent items. This study explores topic modeling using latent Dirichlet allocation (LDA) as a means of identifying enemy items in an operational certification/licensure item bank.

Using Markov Decision Process to Assemble Optimized Testlet Database with Constraints
Jiahe Qian, Educational Testing Service

For the testlet-based assessments that consist of item-blocks, Markov decision process is used to augment a balanced testlet database with constraints. Constraints are imposed to avoid overusing the same blocks. Bellman equations are applied in the process to optimizing the psychometric properties of the testlets included in the database.

Shadow Test Assembly with an Information Target
Jiahui Zhang, Michigan State University; Benjamin Andrews, ACT; Xin Li, ACT

A shadow test assembly method that uses an information target is proposed. A potential advantage is equal measurement precision across the ability scale. Simulation studies showed the proposed method provided more stable precision throughout portions of the ability scale and had more balanced pool usage than the information maximization approach.

Impact of Enemy Items and Repeat-Test Masking on Computerized Adaptive Testing
Timothy Muckle, National Board of Certification and Recertification for Nurse Anesthetists; Kirk Becker, Pearson VUE; Hao Song, National Board of Certification and Recertification for Nurse Anesthetists

Enemy items and repeat tests create challenges in testing operations, especially in a context of continuous administration. This study uses simulation to evaluate how these two factors affect ability estimation and item exposure in computerized adaptive testing. The study finds enemy item relationships helps balance item exposures without impacting precision.

Monday, April 16, 2018
12:25-1:55pm, Manhattan, Individual Presentations, K8

New Directions in Detecting DIF

Session Discussant: Tia Fecter, Pacific Metrics

A Nonparametric DIF Method for Small Focal Groups

Anne Corinne Huggins-Manley, University of Florida; Jingyi He, University of Florida

The purpose of this study is to develop a nonparametric DIF method that (a) allows practitioners to explore for DIF related to small focal groups of examinees, and (b) compares the focal group directly to the composite group that will be used to develop the reported test score scale.

DIF for Accommodated Students with Disabilities: Effect of Differences in Proficiency Distributions

Sarah Quesen, Pearson; Suzanne Lane, University of Pittsburgh

To examine DIF for SWDs, similar *vs*. dissimilar proficiency distributions for the reference group were used. The DIF-free-then-DIF strategy was used with Mantel Haenszel, Wald-1, logistic regression, and HGLM. For the reference group with a similar distribution DIF was not detected; whereas non-IRT methods flagged items using the dissimilar group.

Examining DIF in the Context of CDMs when Q-matrix is Misspecified

Dubravka Svetina, Indiana University; Yanan Feng, Indiana University; Justin Paulsen, Indiana University; Montserrat Valdivia, Indiana University; Arturo Valdivia, Indiana University; Shenghai Dai, Washington State University; Melissa Lee, Indiana University

Development of cognitive diagnostic models is more advanced than test construction using cognitive diagnosis framework. Nonetheless, assessments for diagnostic purposes ought to include high quality items. We demonstrate via a simulation study that traditional approaches to DIF detection fail to identify problematic items when Q-matrix is misspecified.

Using Hierarchical Logistic Regression to Study DIF Variance

Benjamin Shear, University of Colorado Boulder

Most DIF detection methods estimate a single, average DIF coefficient for each item. This paper describes the use of hierarchical logistic regression to test for heterogeneity of DIF in a single item across test contexts, using Monte Carlo simulations and a real data analysis to evaluate the method.

Detecting DIF in Testlet Items: A Polytomous Multilevel Measurement Model vs. Poly-SIBTEST

Wei Xu, National Council of State Boards of Nursing (NCSBN); David Miller, University of Florida

In this study, we proposed a polytomous multilevel measurement model (PMMM-2) and compared it with the poly-SIBTEST in DIF detection for testlet-based polytomously scored items . Researchers and practitioners might consider adopting this model to assist item development and test construction.

Monday, April 16, 2018
12:25-1:55pm, Ambassador II, Coordinated Sessions, K9

Examining standard errors for NAEP group-score comparisons across years and digital transition

Session Chair: Lauren Harrell
Session Discussant: Lauren Harrell

The National Assessment of Educational Progress (NAEP) has begun the transition to digitally-based assessments (DBA), starting with Grades 4 and 8 Mathematics and Reading in 2017. One of the key priorities of the NAEP program is to compare student subgroup performance over time, and the goal of the digital transition is to place the 2017 DBA results onto the trend reporting scale. The 2017 design randomized students to receive either a DBA or pencil-and-paper (PBA) assessment. In order to compare student subgroup results from the 2017 DBA to previous PBA results, the NAEP program has been evaluating the most appropriate methodology for standard error estimation given the transition study design. Specifically, the standard error estimation procedure is reconsidered for population and subpopulation comparisons (a) across multiple years within the same mode, (b) between DBA and PBA modes within a year, and (c) across modes and years. The papers in this session are proposed approaches for either examining the stability of long-term trends, evaluating the digital transition through subgroup differences, or reporting the subgroup results of the DBA assessment on the existing PBA scale while accounting for linking error.

Comparing pairwise chained linking and multi-year concurrent calibration linking in NAEP
Xiaying Zheng, University of Maryland, College Park; Young Yee Kim, American Institutes for Research; Markus Broer, American Institutes for Research; Lauren Harrell, National Center for Education Statistics

Error variances for comparing sub-population scores when linking through random groups design
Xueli Xu, Educational Testing Service; Yue Jia, Educational Testing Service

Jackknife-Based Estimation of Group-Score Standard Errors Incorporating Digital-Based to Paper-Based Linking Error
John Mazzeo, Educational Testing Service; John Donoghue, Educational Testing Service; Bingchen Liu, Educational Testing Service; Xueli Xu, Educational Testing Service

Linking assessments in the presence of a nuisance dimension
Matthew Johnson, Teachers College, Columbia University; Sandip Sinharay, Educational Testing Service

Monday, April 16, 2018
12:25-1:55pm, Gershwin I, Individual Presentations, K10

Electronic Board Session 4

Electronic Board #1

Assessing Dimensionality when Data are Missing Not at Random

Tzu-Chun Kuo, American Institute for Research; Mahmut Gundogdu, University of California Riverside; Ming Lei, American Institute for Research; Hyesuk Jang, American Institute for Research

Four statistical packages/softwares were compared for assessing dimensionality in the aspect of model comparisons when data are missing not at random. Preliminary results suggested that these procedures selected the correct model when the true test structure was unidimensional. However, they favored the multidimensional model when the real structure was bi-factor.

Electronic Board #2

Creating Achievement Level Descriptors with Subject Matter Experts in an Online Environment

Claudia Guerere, Schroeder Measurement Technologies; Angelica Rankin, PLTW

Achievement Level Descriptors (ALDs) are typically conducted in-person. Limitations to in-person meetings include lack of participation from key Subject Matter Experts (SMEs) and cost. This study presents how a successful ALD meeting was conducted in a virtual environment, reducing the cost to produce ALDs and increasing access to SME participation.

Electronic Board #3

Measurement Properties of the College Freshmen Academic Performance Index

Meaghan McMurran, University of California Riverside; Gregory Palardy, University of California Riverside

This study introduces the college *freshmen academic performance index* and examines its measurement properties using CFA and measurement invariance methods. Results indicate a good model fit and consistency across four underrepresented college student groupings (SES, ethnic minority, gender, and ELL), suggesting the index is appropriate for assessing student performance.

Electronic Board #4

Exploring the Accuracy of MIRT Scale Linking Procedures for Mixed-format Tests

Zhen Li, Government of Newfoundland and Labrador; Tianli Li, ACT; Haiqin Chen, American dental association

This study investigates the accuracy of extended Stocking-Lord scale linking procedures for MIRT with common-item nonequivalent-group design for mixed-format tests. Two anchor scenarios are investigated under different levels of format effects (FEs). Results provide recommendations on the appropriateness of UIRT and three MIRT models when FEs presents.

Electronic Board #5

Using Functional Data Analysis to Model Person Response Functions

Kyle Turner, The University of Georgia; George Engelhard, The University of Georgia

This study describes, and considers the benefits of, an approach for estimating person response functions using functional data analysis (FDA). The conceptual contributions are illustrated with data collected on the home environments of preschool children (N=40). The results suggest that FDA offers insights about psychometric issues related to person measurement.

Electronic Board #6

Detecting the Effects of Item Parameter Estimation Methods on Simple-Structure MIRT Equating

Ye Ma, the University of Iowa; Laurentius Susadya, the University of Iowa; Stella Kim, the University of Iowa

The current study aims to detect the effects of using different item parameter estimation methods on Simple Structure MIRT true-score and observed-score equating (Lee & Brossman, 2012; Kim & Lee, 2016) under random group design considering multiple sources of dimensionality, including mixed-format tests and a test with multiple content areas.

Electronic Board #7

Improving Students' Workforce Readiness Preparation Using O*Net: A Focus on Communication Skills

Maria Elena Oliveri, Educational Testing Service; Rene Lawless, Educational Testing Service; Laura McCulla, Educational Testing Service; Jonathan Schmidgall, Educational Testing Service

We discuss workplace readiness preparation to better align workplace skills with instruction and assessment. We discuss results of our O*Net analysis to identify critical communicative activities relevant across job zones and industries. We illustrate an evidence-centered design approach to identify and design workplace English communicative tasks.

Electronic Board #8

Dynamic IRT analysis in Adaptive Learning Systems

Jung Yeon Park, University of Leuven

Student monitoring systems requires to follow-up learning progress of individual students to provide tailored instructions. Despite a rapid improvement in the dynamic parameter estimation, some issues still remain. In this study, we develop and propose an advanced Elo rating system that can address several practical challenges.

Electronic Board #9

Evaluating Instruction Efficacy with Logistic Regression and Discriminant Analysis

Logan Rome, Curriculum Associates

This study utilizes logistic regression and discriminant analysis to predict meeting yearly growth targets from time on task and pass rate for an online instructional tool. The efficacy of the online instruction will be evaluated and classification decisions for the two prediction methods will be compared.

Electronic Board #10

Testing the Orthogonal Assumption of the Bifactor IRT Model: A Bayesian Approach

Ken Fujimoto, Loyola University Chicago

Item response theory (IRT) models that specify a bifactor structure for the data assume all dimensions are orthogonal to each other. A Bayesian multilevel multidimensional IRT model is used to examine how the item parameter estimates and measurement reliabilities are impacted from assuming orthogonality when the data violates this assumption.

Electronic Board #11

Investigating the Impact of Pool Size and Item Characteristics on CAT

Yi He, ACT; Ann Wang, ACT; Stephanie Su, ACT

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

Electronic Board #12

Obtain Growth Expectations based on Conditional Distribution and Regression

Ping Yin, Curriculum Associates

This study evaluates the feasibility and effectiveness of two approaches (conditional distribution and regression) to obtain growth expectations. These approaches are investigated using data from a computer adaptive assessment administered multiple times a year. The amount of prediction error associated with these approaches in modeling growth is evaluated through cross-validation.

Electronic Board #13

Modeling Item and Block Response Time Distributions in a NAEP Mathematics Test

Jessica Feng, Educational Testing Service; Bingchen Liu, Educational Testing Service; Gary Feng, Educational Testing Service

Using item and block timing data from a 2016 NAEP mathematics pilot study, we investigated how certain item and student characteristics (item type, item position, item difficulty, and student overall speed) affect item response time and total block time distributions.

Electronic Board #14

Comparing math achievement by domains for high school students and high-school-equivalency population

Lida Chen, The University of Iowa; Catherine Welch, The University of Iowa; Stephen Dunbar, The University of Iowa; Timothy Davey, Educational Testing Service

This study compares the performance of traditional high school students and with examinees that are completing a high school equivalency assessment in different math domains. IRT-based methods are used to estimate the ability distributions for both samples of examinees by domain. The results may imply test development and preparation.

Electronic Board #15

Optimal Scripted On-the-Fly Multistage Tests with Passages

Xiao Li, University of Illinois at Urbana-Champaign; Bruce Williams, ACT; Sung-Hyuck Lee, ACT; Hua-Hua Chang, University of Illinois at Urbana-Champaign

Scripted On-the-fly Multistage Testing (SOMST) borrows merits from computer adaptive tests (CAT) and multistage tests (MST), and is as easily implemented with passages as with discrete items. This paper investigated several optimal designs of SOMST which approach CAT reliability while maintaining good item usage and low item overexposure.

Electronic Board #16

Exploring the Measurement of Collaborative Problem Solving Using a Human-Agent Educational Game

Steve Polyak, ACTNext ACT Inc; Kristin Stoeffler, ACT Inc

Electronic Board #17

Interaction and Sub-Skills Scoring Methods for Collaborative Problem Solving Human-Agent Assessment

Pravin Chopade, ACTNext ACT Inc; Alina von Davier, ACTNext ACT Inc.; Yigal Rosen, Harvard University

Electronic Board #18

Collaborative Problem Solving Human-Agent Assessment and the Big Five

Samuel Greiff, University of Luxemburg; Katharina Herborn, University of Luxemburg; Maida Mustafic, University of Luxemburg

Electronic Board #19

CPS-evaluator: An automated assessment of collaborative discourse

Jiangang Hao, Educational Testing Service; Lei Chen, Educational Testing Service; Michael Flor, Educational Testing Service; Lei Liu, Educational Testing Service; Alina von Davier, ACTNext ACT Inc; Jessica Andrews, Educational Testing Service; Patrick Kyllonen, Educational Testing Service

Monday, April 16, 2018
2:15-3:45pm, Majestic II, Coordinated Sessions, L1

Testing in the professions: Credentialing policies and practice

Session Chair: Susan Davis-Becker, ACS Ventures, LLC
Session Discussant: Wayne Camara, ACT

Psychometrics and practices in standardized testing can be found in many contexts and although the purpose and types of examinees may be different, the guiding principles are consistent (e.g., *Standards for Educational and Psychological Testing*). As practitioners, the practice of applying these principles may vary between contexts such as education and credentialing. *Testing in the Professions*, a recently published volume in the *NCME Applications of Educational Measurement and Assessment* series, provides a unique and comprehensive review of the current practices in credentialing testing (e.g., certification, licensure) covering of program/test design, development, and use. In this session, the editors and several of the contributing authors will review the current state of credentialing practices (e.g., program purpose and use), how these programs are designed from top to bottom (including key decisions to be made at the outset), approaches as to how the content of credentialing programs can be defined, how test scores are estimated, interpreted and maintained for programs of various sizes, and how the technical quality and integrity of these programs are evaluated by external entities.

Credentialing: A continuum of measurement theories, policies, and practices
Chad Buckendahl, ACS Ventures, LLC

Test Design: Laying out the Roadmap
Timothy Muckle, NBCRNA; Susan Davis-Becker, ACS Ventures, LLC

Specifying the Content of Credentialing Examinations
Amanda Clauser, National Board of Medical Examiners; Mark Raymond, National Board of Medical Examiners

Estimating, Interpreting, and Maintaining the Meaning of Test Scores
Kathleen Gialluca, Pearson VUE; Walter (Denny) Way, The College Board

Using Standards to Evaluate Credentialing Programs
Larry Fabrey, PSI Services

Monday, April 16, 2018
2:15-3:45pm, Broadway I, Coordinated Sessions, L2

Improving Human Rating

Session Chair: Edward Wolfe, Educational Testing Service
Session Discussant: Robert Johnson, University of South Carolina

This coordinated paper session of four papers and a discussant focuses on applied research that seeks to improve the human scoring enterprise by identifying potential inefficiencies or ineffectiveness in current practices, proposing alternatives to those practices, and empirically evaluating the relative effectiveness and efficiency of the proposed alternatives. Paper 1 (Wolfe) focuses on the impact that practice applying a scoring rubric during rater training has on subsequent rater performance during certification and operational scoring. Paper 2 (Attali) investigates the impact of three approaches to rater feedback on rater accuracy. Paper 3 (Cohen) compares several methods of calibrating (equating) rating data. Paper 4 (Black & Meadows) reports on research that seeks to disentangle rater judgment error from legitimate differences of opinion during appeals processes that are common in examinations administered in the United Kingdom. Robert Johnson of the University of South Carolina will serve as the Discussant for the session. Jointly, these four papers examine a range of potential improvements across the entire duration of the human rating process.

Impact of Extended Practice during Rater Training on Rater Accuracy and Speed
Edward Wolfe, Educational Testing Service; Ikkyu Choi, Educational Testing Service; Larry Davis, Educational Testing Service; Nancy Glazer, Educational Testing Service; Cathy Wendler, Educational Testing Service

Effect of Immediate Feedback and Type of Feedback on Rater Accuracy
Yigal Attali, Educational Testing Service

Comparing the Accuracy of Methods for Equating of Rating Scales
Yoav Cohen, NITE

Rater Error versus Tolerable Uncertainty
Beth Black, OfQual; Michelle Meadows, OfQual

Monday, April 16, 2018
2:15-3:45pm, Broadway II, Coordinated Sessions, L3

Within and between-high school measurement challenges in college admission

Session Chair: John Hansen, Harvard University
Session Chair: Daniel Koretz, Harvard University
Session Discussant: Matthew Gaertner, SRI International

Predictive validity studies have led to a consensus among researchers that a combination of high school grades and standardized test scores predicts college grades better than either variable independently. One hypothesis for the unique predictive power of high school GPA and standardized test scores in OLS regression models is that high school GPA is scaled within schools, while standardized test scores are on a common scale across all schools (Zwick & Green, 2007; Kostal, Sackett, & Kuncell, 2017). This session presents research that tests and explores this hypothesis, along with its equity implications for college admission policy. The first paper shows that trends in high school grades and SAT scores have diverged in recent decades. The second paper finds that relying on class rank for college admission—a potential solution for addressing between-school variability in grading—could amplify bias attributable to inaccurate GPA adjustments for course difficulty. The third paper uses multi-level modeling to estimate the role of high school characteristics in differential prediction for black and Hispanic students. The fourth paper seeks to improve upon previous efforts to create an index of high school academic rigor by accounting for differences across high schools in course difficulty.

High School Grade Inflation: 1998 to 2016
Michael Hurwitz, College Board; Jason Lee, University of Georgia

Using a Graded Response Model to Analyze HSGPA Weighting Policies
John Hansen, Harvard University; Philip Sadler, Harvard-Smithsonian Center for Astrophysics; Gerhard Sonnert, Harvard-Smithsonian Center for Astrophysics

Advances in an Empirically-Derived Index of High School Academic Rigor
Jeffrey Allen, ACT, Inc.; Krista Mattern, ACT, Inc.; Dina Bassiri, ACT, Inc.

The Role of School Characteristics in Differential Prediction for Disadvantaged Students
Preeya Mbekeani, Harvard University; Daniel Koretz, Harvard University

Monday, April 16, 2018
2:15-3:45pm, Broadway III, Coordinated Sessions, L4

Exploring the Potential Impact of SEL Assessment on School Practices

Session Chair: Kevin Petway, The Enrollment Management Assocation
Session Chair: Jinghua Liu, The Enrollment Management Assocation
Session Discussant: Patrick Kyllonen, Educational Testing Service

A plethora of research suggests that social and emotional learning (SEL) skills are critical for student success in school, work, and life generally. While there are plenty of assessments that have been developed to measure these skills, it is often less clear how to apply information from assessments to school practices. This symposium explores how several assessments can be or are used in schools to better understand students and improve student outcomes. Two first studies present the results of research conducted to better understand the assessments themselves, and describe how data from the assessments can be used to drive positive changes in students. The third study discusses how a set of schools are currently using data from a longitudinal assessment of SEL to improve student development. The final paper shifts to address how a large number of schools are using data from an SEL assessment to make admissions decisions.

Assessing, interpreting, and encouraging SEL skills development among high school students
Sam Rikoon, Educational Testing Service

Differences in the development of essential SEL skills across high school curriculum
Ross Anderson, Education Policy Improvement Center; Paul Beach, Educational Policy Improvement Center

From data to practice: The Mission Skills Assessment in action
Meghan Brenneman, The Enrollment Management Assocation

The use of an SEL assessment for middle and high school admissions
Kevin Petway, The Enrollment Management Assocation

Monday, April 16, 2018
2:15-3:45pm, Ambassador III, Individual Presentations, L5

IRT for Next Generation Assessments

Session Discussant: Quinn Lathrop, Pearson

IRT-Based Simulation Study in-Context: Comparing Two Calibration Models for Next-Generation Assessments
Nurliyana Bukhari, Universiti Utara Malaysia

Using a computer simulation study based on next-generation assessments, I employed Luecht and Ackerman's (2017) expected-response-function-based residuals approach to evaluate the performance of the UIRT and MIRT calibration models. I found that the MIRT model tends to produce less estimation error and fit the data better than the UIRT model.

Understanding Student Performance in Contextualized Science Tests: A Cognitive Diagnostic Model Approach
Dongsheng Dong, University of Washington; Min Li, University of Washington; Jim Minstrell, Facet Innovations; Maria Araceli Ruiz-Primo, University of Standford

This paper applies the GDINA model to examine students' mastery of nine attributes in a middle school physical science test. The goals of this study are to provide diagnostic information about students' understanding of Force and Motion topic and to investigate the impact of two context-level features on students' performance.

An Item Response Theory Model for Next Generation of Science Standards Assessments
Frank Rijmen, AIR; Ahmet Turhan, AIR; Tao Jiang, AIR

An IRT approach is proposed for item clusters that assess the Next Generation of Science Standards. The model takes local dependencies into account by incorporating nuisance dimensions. Proficiency estimates are obtained by maximizing the likelihood after marginalizing out the nuisance dimensions.

Jointly Scaling a General Assessment with On Demand Assessments of Individual Standards
Nathan Dadey, The National Center for the Improvement of Educational Assessment, Inc.; Shuqin Tao, Curriculum Associates; Jennifer Dunn, The National Center for the Improvement of Educational Assessment, Inc.

This work examines whether two types of interim assessments – a "general" assessment that broadly covers grade-level content standards and a set of twenty nine short "mini-assessments" that cover individual standards or sub-standards – can be placed onto a single unidimensional reporting scale.

Monday, April 16, 2018
2:15-3:45pm, Majestic I, Individual Presentations, L6

Impact of People on Linking and Equating

Session Discussant: Jennifer Dunn, Questar Assessments

Linking CR Scores using MC Scores with MDIA Weighting in Small Samples
Yanmei Li, Educational Testing Service

Recently, a linking method for test forms without anchor items using minimum discriminant information adjustment (MDIA) was proposed (Haberman, 2015). In this study, the accuracy of this linking method in small samples was investigated using real data sets from an English language test.

Equating at the Passing Score for Credentialing Exams with Small Sample Sizes
Amanda Wolkowitz, Alpine Testing Solutions; Keith Wright, The Enrollment Management Association

Our study investigates multiple methods under the common item, non-equivalent groups design for effectively equating at a single pass/fail score using small sample sizes versus equating scores across an entire total score scale. Using our results from both real and simulated data, we discuss the practical implications of the results.

Linking HElghten Critical Thinking Scores across Forms in US and International Samples
Zhen Wang, Educational Testing Service; Usama Ali, Educational Testing Service; Joseph Rios, Educational Testing Service; Guangming Ling, Educational Testing Service; Yu Sun, Educational Testing Service

Several college level learning outcome assessments designed in US were adopted by foreign institutions for the purpose of assessing learning outcome. The global students may be different from the original US sample. We gather evidence to support the appropriateness of test score linking and reporting for the international sample.

Evaluating Group Equivalence in a Random Groups Design
Michael Walker, The College Board; Sooyeon Kim, Educational Testing Service; Timothy Moses, The College Board

This paper explores methods for testing the equivalence of groups in the random groups equating design. Both proposed methods combine information on pretest item statistics with observed performance on test items. The paper provides guidelines for using the methods and suggested remedies for group nonequivalence.

Monday, April 16, 2018
2:15-3:45pm, Plymouth, Individual Presentations, L7

Learning Progressions: Development and Evaluation

Session Discussant: Anna Topczewski, GED Testing Service

Understanding Learning Progression of Students with Cognitive Disabilities Using Performance Level Descriptors

Luxi Feng, Texas A&M University-College Station; Chris Domaleski, National Center for the Improvement of Educational Assessment

The purpose of this study is to describe the expertise of students with significant cognitive disabilities (SCD) using performance level descriptors (PLD) of prominent alternate assessments based on alternate achievement standards (AA-AAS). We analyzed the PLD, summarizing key dimensions to understand underlying learning progressions and inform the measurement of growth.

Validating a Vertical Scale based on Learning Progressions

Ping Yin, Curriculum Associates; Marie Tranguch, Curriculum Associates

Learning progressions describe how learning occurs and advances. Because LPs are primarily developed from professional judgement of experts, it is important to validate whether such theoretical expectations can be supported empirically. This study validates a vertical scale based on LPs developed from curricular sequencing and instructional activities.

Developing a Learning Progression for the Crosscutting Concept of Energy

Rajendra Chattergoon, University of Colorado Boulder; Derek Briggs, University of Colorado Boulder; Borbala Mahr, University of Colorado Boulder; Erin Furtak, University of Colorado Boulder

This paper explores how the NGSS crosscutting concept of Energy and Matter can be represented and modeled psychometrically as across- and within-discipline learning progressions. This study uses data from 65 items given to 939 high-school students. Preliminary results are mixed but lend support for an across-discipline energy learning progression.

The Interpretation of Learning Progressions: Do Teachers and other Subject-Matter Experts Agree?

Edith Graf, Educational Testing Service; Peter van Rijn, Educational Testing Service Global

A learning progression (LP) is useful if teachers can use it to effectively guide instruction. Although research on the empirical validation of LPs exists, little is known about how teachers interpret them. We explored agreement among teachers and between teachers and other experts on the interpretation of two different LPs.

Evaluating a Learning Progression Theory: Comparative results from two psychometric models.

Duy Pham, University of Massachusetts Amherst; Malcolm Bauer, Educational Testing Services; Caroline Wylie, Educational Testing Services; Craig Wells, University of Massachusetts Amherst

We used higher-order sequential cognitive diagnosis models to evaluate a theory underlying two research-based learning progressions for middle-school algebra. Latent class estimates provided some convergent evidence with a prior study that used multi-dimensional item response theory. Educational implications are discussed.

Monday, April 16, 2018
2:15-3:45pm, Manhattan, Individual Presentations, L8

Reliability of Scores and Subscores

Session Discussant: Usama Ali, Educational Testing Service

Length of subscores and reliability of diagnostic information
Samuel Livingston, Educational Testing Service; Omar Santos-Cedeno, Educational Testing Service

Diagnostic information provided by subscores is useful only to the extent that it is consistent across test forms. By restructuring data from a long test with three subscores, we created alternate-forms data to investigate the consistency, across forms, of differences between subscores 21, 14, and 7 items long.

Comparison of NAEP scale reliabilities in high and low performing populations
Andrew Kolstad, P20 Strategies LLC

An empirical Bayes model undergirds a newly developed reliability measure, which is based on a model of posterior distributions and their sampled plausible values. The reliability of NAEP eighth-grade scales is compared in 2013 mathematics and reading and 2011 science and writing nationally and in high- and low-performing student populations.

Item-score reliability for educational tests
Eva Zijlmans, Tilburg University; Jesper Tijmstra, Tilburg University; L. Andries Van der Ark, University of Amsterdam; Klaas Sijtsma, Tilburg University

When constructing an educational test, item-score reliability is a useful tool to investigate the repeatability of item scores. In this study, the relationship between item-score reliability and other item indices is investigated by means of a simulation study. Educational data will be used to provide an empirical example.

Can We Learn from the Past?: Using Previous Scores to Augment Subscores
Whitney Smiley, American Board of Internal Medicine

Traditionally, data augmentation of subscores is completed by taking ancillary information within the *same* test to augment subscores. This research examines whether augmenting subscores using ancillary information *across* one and two testing occasions has potential to stabilize subscores.

Using Simulation to Evaluate Retest Reliability of Assessment Results
Brooke Nash, University of Kansas; Amy Clark, University of Kansas; William Thompson, University of Kansas

As diagnostic assessment systems become more prevalent, alternatives to traditional reliability methods must be explored. One method for evaluating retest reliability when practical constraints make a second empirical measurement infeasible is simulation methodology. This paper summarizes an application of using simulation to report reliability for one operational assessment program.

Monday, April 16, 2018
2:15-3:45pm, Ambassador II, Coordinated Sessions, L9

Evaluating Paper and Computer Adaptive Test Score Comparability from Multiple Perspectives

Session Chair: Laurie Davis, ACT
Session Discussant: Leslie Keng, Center for Assessment

There is an ongoing movement towards computerized assessments to facilitate administration, scoring, and reporting processes. However, paper-based assessments still make up a large proportion of standardized test administrations in the United States. Addressing comparability is complicated but necessary for programs that provide scores based on more than one type of administration. This coordinated paper session adds to the body of research specifically looking at comparability issues arising from paper and computerized adaptive testing (CAT) administrations. Comparability is evaluated from a variety of perspectives including construct equivalence, scale score, and measurement error, as well as investigating the impact of various equating methods and timing conditions on score comparability. The studies include both empirical data from a study where students were randomly assigned to paper and CAT conditions, as well as simulation results. Statistical adjustments are described to address non-comparable results. Additionally, the tension between paper and CAT comparability and the goals of CAT (e.g., reduced measurement error, shorter test lengths) is discussed.

Comparability of Constructs for Paper and Computer Adaptive Tests
Shalini Kapoor, ACT; Yi-Fang Wu, ACT

Comparability of Speededness for Paper and Computer Adaptive Tests
Hongwook Suh, ACT; Yi-Fang Wu, ACT; Benjamin Andrews, ACT; Sonya Powers, ACT

Comparability of Scale Scores for Paper and Computer Adaptive Tests
Youngwoo Cho, ACT; Yang Lu, ACT; Tianli Li, ACT; Qing Yi, ACT

Comparability of Paper and Computer Adaptive Test Scores under Different Equating Frameworks
Benjamin Andrews, ACT

Comparability of Measurement Precision for Paper and Computer Adaptive Tests
Yang Lu, ACT; Qing Yi, ACT; Yi He, ACT; Tony Thompson, ACT

Monday, April 16, 2018
2:15-3:45pm, Gershwin I, Individual Presentations, L10

Electronic Board Session 5

Electronic Board #1

Applying the Continuous Beta Response and Beta Unfolding Models to Response Time

Weldon Smith, University of Nebraska Lincoln, Buros Center for Testing; HyeSun Lee, California State University Channel Islands

A continuous beta response and beta unfolding model were applied to response time data. The unfolding model offers a new way to understand response times for both items and individuals, modeling items as having an ideal response time and scoring individuals based on their proximity to that time.

Electronic Board #2

Factor Mixture Analysis of a Large Scale Hybrid Assessment

HyunJoo Jung, University of Massachusetts Amherst; Pamela Kaliski, The College Board; Lei Wan, The College Board

Person-centered approaches such as factor mixture modeling and latent class modeling can improve the utility of performance assessment data in education by identifying groups of students who have similar profiles. We investigate latent classes of a large scale hybrid assessment which comprises two through-course components and one end-of-course exam.

Electronic Board #3

Machine Learning based Item Response Prediction for Mixed-Format Tests

Shumin Jing, University of Iowa; Sheng Li, Adobe Research

The purpose of this study is to design a machine learning approach to predict the item response in mixed-format tests. Particularly, a collaborative filtering model is used to extract latent factors for examinees and test items. Empirical results from a simulation study validate the effectiveness of the proposed method.

Electronic Board #4

Categorical or Dimensional Models for Cognitive Processes of Children's Mathematical Abilities

Yi-Ling Cheng, Michigan State University; V. Rani Satyam, Michigan State University; Mark Reckase, Michigan State University

Previous research has debated whether psychological constructs are categorical or dimensional. The representation of categorical or dimensional might change depending on the place in the learning process. The presented study conducted model comparisons on a range of mathematics performances from TIMSS 2011 to test the hypotheses.

Electronic Board #5

Detecting Multidimensional Differential Item Functioning

Ming Lei, American Institutes of Research; Okan Bulut, University of Alberta; Hyesuk Jang, American Institutes for Research

The study investigate multidimensional DIF using real data and simulations. The differences between the mean differences of domain distributions and the magnitude of correlations among domains are considered. Methods included are the generalized multi–group bifactor DIF model, the bifactor models with and without constraints, and the logistic regression approach.

Electronic Board #6

Estimating Standard Error of Equating for the Non-Equivalent Anchor Test Design

Caiyan Zhang, The College Board; Judit Antal, The College Board

Chained equipercentile method (CEM) and non-equivalent anchor test (NEAT) equating design are viable equating method and design that testing programs use. And yet, performance of CEM under NEAT design has not been studied widely. This study fills up the gap and will provide practical guidance for equating practitioners.

Electronic Board #7

Long-term English Learners' Performance on English Language Proficiency Tests and Content Assessments

Nami Shin, University of California, Los Angeles

This study examines the relationship between long-term English Learners' (LTELs) performance on English Language Proficiency (ELP) tests and their performance on content assessments. Analyzing longitudinal student-level data from a large urban school district, this study shows LTELs' performance trajectories in the assessments and the relationship between the two tests.

Electronic Board #8

The Comparison of Reliability Estimates in Multidimensional Tests

Xiaolin Wang, University of Kansas; Shenghai Dai, Washington State University; Justin Paulsen, Indiana University; Ou Zhang, Pearson

The current study aims to examine and compare the performance of Cronbach's alpha and three estimators designed for multidimensional tests (Stratified alpha, maximal reliability, and generalized McDonald's omega) across combinations of test dimensionality, dimension length, sample size, correlation between dimensions, and test structures.

Electronic Board #9

Instructional Equivalence via Distractor Analysis: Examining International Assessments

John Poggio, University of Kansas

International assessments explore country standing, and magnitude of differentiation among countries. This investigation studies whether examinees in different countries make comparable errors regardless of scores attained. Analyses show that errors within countries are similar regardless of scores attained, but error analyses reveal across-country differences suggesting country instruction is not equivalent.

Electronic Board #10

Applying the Projected IRT Model to Correct for Inconsistent Score Scale Interpretation

Terry Ackerman, ACT; Edward Ip, Wake Forest School of Medicine; Shyh-Huei Chen, Wake Forest School of Medicine; Tyler Strachan, university of North Carolina at Greensboro; Yanyan Fu, University of North Carolina at Greensboro; John Willse, university of North Carolina at Greensboro

Test items often maintain a broad content to be ecologically valid. It is not uncommon to have difficult items disproportionally require additional abilities resulting in a confounding of difficulty and dimensionality - a phenomenon we call inconsistent score scale interpretation. This study highlights a novel approach to overcome this phenomenon.

Electronic Board #11

Preliminary Validity Evaluation of a Learning Progression for the Concept of Function

Stephanie Peters, Educational Testing Service; Edith Graf, Educational Testing Service; James Fife, Educational Testing Service

Learning progressions are working models of cognitive development that may require revision in light of empirical evidence. The goal of the current study is to advance the validation of a learning progression for the concept of function. Connections between assessment design, teacher instruction, and student learning will be discussed.

Electronic Board #12

Effects of First Impressions on Teachers' Ratings of Student Behavior

Sattik Ghosh, UC Davis; Megan Welsh, Adviser/Co-Author; Sandra Chafouleas, Researcher; Greg Fabiano, Researcher; T.C. Riley-Tillman, Researcher

The current study utilizes data from a study of a behavioral rating scale to explore the effects of first impressions on student behavior ratings. Preliminary results indicate that black male and black female students receive lower average Respectfulness ratings if they exhibit slightly disruptive behavior on a rater's first viewing.

Electronic Board #13

Beta True Score Equating for the Common-Item Nonequivalent Groups Design

Shichao Wang, ACT, Inc; Won-Chan Lee, The University of Iowa; Michael Kolen, The University of Iowa

This study aims to gain a better understanding of the factors that affect the accuracy of beta true score equating for the common-item nonequivalent groups design. A variety of simulation conditions are included in the design, including the sample size, group difference, and equating methods.

Electronic Board #14

Developing a Cognitive Diagnostic Assessment to Measure Misconceptions in Newton's Laws

Mary Norris, Virginia Tech

Few CDAs have been developed specifically for CDMs and there is little information in the literature on methods for doing so. This research reports on the process of developing a CDA to measure misconceptions in physics using Bradshaw and Templin's (2014) Scaling Individuals and Classifying Misconceptions (SICM) model.

Electronic Board #15

Determining the Cut Score on a Universal Screener Test with Case-Control Sampling

Xin Luo, Uber

Setting accurate cut scores on a universal screener test has significant consequences for students. However, this process may be complicated by the skewed classifications. This study proposed a new sampling method and verified its implementation by comparing it with the traditional sampling in various test settings.

Electronic Board #16

The cognitive diagnosis analysis of reading comprehension

Yaping Liu, Beijing Normal University; HongB166o Wen, Beijing Normal University; Faming Wang, Beijing Normal University

This study aim to compare the performance of five commonly used cognitive diagnosis models with the Chinese reading comprehension and explore the effects of attribute granularity to CDA. The result illustrate that G-DINA is best, followed by RRUM. The effects of attribute granularity to CDA be discussed.

Electronic Board #17

Scoring for Incomplete Computerized Adaptive Tests

Ching-Wei Shin, Pearson

The purpose of this paper is to propose a method (the modified penalty scoring method) for scoring incomplete computerized adaptive tests and to investigate the impact of applying this method. The methodology and results from the paper will provide valuable guidance and reference to scoring incomplete tests in CAT programs.

Electronic Board #18

Comparison of Short-length Testlet-based CAT and MST under Rasch Testlet Models

Seohong Pak, National Board of Medical Examiners; Catherine Welch, University of Iowa; Stephen Dunbar, University of Iowa

This simulation study (56 conditions) was designed to investigate the impacts of total test lengths, testlet random effect sizes, and ability estimation methods on the measurement accuracy for a short-length CAT and MST comprised only of testlsts.

Electronic Board #19

Comparison of concurrent and separate MIRT linking of domain and overall scores

Moonsoo Lee, Korea Institute for Curriculum and Evaluation(KICE)

The purpose of this research is to compare the performance of concurrent calibration and several separate MIRT linking methods for both domain and overall scores. The results of this study suggest that concurrent calibration generally performs better than separate linking methods for domain and overall scores when groups are non-equivalent.

Electronic Board #20

Comparing two Estimation Algorithms for Mixture Rasch Models using R packages

Yevgeniy Ptukhin, Southern Illinois University Carbondale; Yanyan Sheng, Southern Illinois University Carbondale

Mixture Rasch models can be estimated using conditional maximum likelihood (CML) or joint maximum likelihood (JML) methods, which are compared in this study using Monte Carlo simulations. The results indicate that JML is preferred with BIC in identifying the correct number of classes and CML is preferred in parameter recovery.

Monday, April 16, 2018
4:05-6:05pm, Majestic II, Coordinated Sessions, M1

Procedures for Detecting Aberrant Exam-Taking Behavior in the Operational Setting

Session Chair: Huijuan Meng, Graduate Management Admission Council
Session Chair: James Wollack, University of Wisconsin–Madison

Cheating damages the integrity of a testing program and can cause testing organizations significant losses. Security breaches can arise from individuals memorizing and sharing items, the concerted efforts of a test preparation company to harvest items and teach them to their customers, and answer copying or collusion among examinees during a testing event. Without proper detection, these types of cheating could remain undetected until their presence becomes significant enough to threaten test-score validity. Therefore, effectively identifying cheaters is a popular topic in the measurement field. Many detection techniques have been developed to flag aberrant testing behaviors. Some of them are specifically designed for the paper-and-pencil test and may not be feasible for pool-based computer-administered tests. Some of them are based on complicated mathematical models and extensive ad-hoc data analyses and thus cannot be practically fitted into many testing programs' operational schedules. This session covers several methods that could help detect aberrant testing behavior in the operational setting. Findings in this session may provide more insights into data forensics research. They may also inspire practitioners to use more practical, less time-consumingly computed statistics in their operational work to improve test security.

Detect Compromised Items in a CAT Licensure Exam
Hong Qian, NCSBN; Ada Woo, ACTNext by ACT, Inc.

Revoking Test Scores for Aberrant Records: Are We There Yet?
Huijuan Meng, GMAC

Monday, April 16, 2018
4:05-6:05pm, Broadway I, Coordinated Sessions, M2

Advances in IRT Equating: Old Methods and New Tricks

Session Chair: Tim Moses, The College Board
Session Discussant: Mike Edwards, Arizona State University
Session Discussant: Michael Kolen, The University of Iowa

This proposed session includes five papers that cover advances in applications of Item Response Theory (IRT) to equating. The first paper presents an overview of requirements of equating, introducing challenges in IRT applications that are addressed in the other four papers. The second paper covers equity evaluations using multidimensional IRT models appropriate for multidimensional tests. The third paper addresses the extent to which estimation accuracy of IRT parameters and observed score distributions can be improved through estimation approaches that do not assume normality. The fourth paper considers equating of tests and composites using approaches that account for the multidimensionality of the composite as a simple structure of the tests. The fifth paper addresses the definition of highest and lowest scores of IRT true score equating, comparing arbitrary linear extrapolations to approaches that more closely preserve characteristics of the true score equating function. Two experts in psychometrics will provide final concluding discussions of the papers and presentations.

Introduction to Advances in IRT Equating: Old Methods and New Tricks
Tim Moses, The College Board

Equity Properties of Multidimensional Item Response Theory Equating Methods
Won-Chan Lee, The University of Iowa; Stella Kim, The University of Iowa; Jaime Malatesta, The University of Iowa

Item Response Theory Equating Without Normality Assumption for Estimation Procedures
Hyung Jin Kim, The University of Iowa

Comparison of Three Possible Approaches to Composite-Score Equating
Stella Kim, The University of Iowa; Tim Moses, The College Board

IRT true score equating 2.0
YoungKoung Kim, The College Board; Tim Moses, The College Board

Monday, April 16, 2018
4:05-6:05pm, Broadway II, Coordinated Sessions, M3

Promises and Challenges with Computerized-Adaptive Testing in K-12 Assessments

Session Chair: Liru Zhang, Delaware Department of Education
Session Chair: Ronald Hambleton, University of Massachusetts - Amherst

Computerized Adaptive Testing (CAT) is becoming more common in state assessments for high-stakes accountability. It is anticipated that the implementation of adaptive testing will be steadily escalated in K-12 education, especially for diagnostic testing. Over the course of testing, attractive advantages of adaptive testing, such as high efficiency and greater precision in measurement across ability levels, may provide possible solutions to some issues in large-scale assessments. However, special features in K-12 assessments, such as the high-stakes nature and heavy influence of educational policies, large student populations with diverse backgrounds, wide range of academic achievement levels, persistent achievement gaps, the comparability of test scores from different testing conditions (e.g., accommodations) and across multiple test versions, and the broader content standards and curriculum, multiple-grade measures, and on-grade test content, present tremendous technical challenges for the design, development, and implementation of adaptive testing.

In this symposium, four well-known educational researchers and an experienced moderator provide a structured discussion organized around five categories of CAT: Test Design, Item Pool Development, Psychometric Considerations, Test Delivery and Administration, and Validity Evidence in K-12 assessments. Following the meeting, a transcript including questions and answers will be prepared and disseminated to interested readers.

Promises and Challenges with Computerized-Adaptive Testing in K-12 Assessments
Suzanne Lane, University of Pittsburgh; Richard Luecht, University of North Carolina Greensboro; Matthew Schulz, Smarter Balanced Consortia Assessments; Walter (Denny) Way, College Board

Monday, April 16, 2018
4:05-6:05pm, Broadway III, Coordinated Sessions, M4

Collaborating to Measure Collaboration Skills: Principles, Methodologies, and Lessons Learned

Session Chair: Jessica Andrews, Educational Testing Service
Session Discussant: André Rupp, Educational Testing Service

Collaborative problem solving (CPS) is a critical competency in a variety of educational and professional work contexts. Despite the importance and relevance of CPS skills in many contexts, only more recently has assessment work and curriculum reform begun to focus to a greater extent on the acquisition, development, and assessment of this 21st century skill. Given the complexity of the tasks used for assessing CPS and the associated resulting data structures, it is indispensable to work towards scaling up resulting assessment solutions based on robust architectures. While there is some published research literature available on conceptual frameworks for CPS, specific tasks that have been designed, or general validity evidence surrounding the assessment, little is often shared about how the interdisciplinary collaboration can be made more effective. In this coordinated session, we narrow this gap between theory and practice. We bring together representatives from four research teams who have engaged in this work to discuss the key design decisions, collaborative processes, associated tools / artifacts, as well as the practical lessons learned from these projects.

The In-task Assessment Framework
Jessica Andrews, Educational Testing Service; Carol Forsyth, Educational Testing Service

NAEP
Julie Coiro, University of Rhode Island; Jesse Sparks, Educational Testing Service; Carita Kiili, University of Oslo; Jill Castek, University of Arizona

PISA
Arthur Graesser, University of Memphis

Assessment & Teaching of 21st Century Skills
Johanna Pöysä-Tarhonen, University of Jyväskylä; Päivi Häkkinen, University of Jyväskylä; Esther Care, University of Melbourne; Nafisa Awwal, University of Melbourne

Monday, April 16, 2018
4:05-6:05pm, Ambassador III, Individual Presentations, M5

Fairness in Testing Policies and Practices

Session Discussant: Maria Elena Oliveri, Educational Testing Service

The Use of Test Scores from Large-scale Assessments: Psychometric and Statistical Considerations
Henry Braun, Boston College; Matthias von Davier, National Board of Medical Examiners

Measures of student achievement from large-scale assessments (e.g. NAEP, PISA) employing plausible values (PV) differ from end-of-course test scores or SAT/ACT scores, with important implications for utilization and interpretation. This paper presents a comprehensive review of the psychometric characteristics of PV and addresses issues regarding their use in secondary analyses.

Improving Validity in Image-Based Assessment Using Simplified Line Drawings
Frank Padellaro, University of Massachusetts Amherst; Darius Taylor, University of Massachusetts Amherst; Lisa Keller, University of Massachusetts

Where applicable, image-based assessments can reduce construct irrelevant information by working across language or cultural barriers (Keller, L., Keller, R., Nering, M., 2017), but little research has focused on improving such items. This study compares the use of simplified line drawings to more detailed photographs and written items.

Investigating test prep impact on score gains using quasi-experimental propensity score matching
Edgar Sanchez, ACT, Inc; Raeal Moore, ACT, Inc.; Maria Ofelia San Pedro, ACT, Inc.

A quasi-experimental method was used to explore the impact of participating in test preparation prior to retesting, whether the impact of test preparation depends on the first ACT score, and the impact of specific test preparation activities on retest scores. Findings are used to explore fairness issues in admissions testing.

Using Propensity Score Matching to Examine How Accommodations Affect Reading Performance
Sarah-Truclinh Tran, NWEA; Xueming Sylvia Li, NWEA; Wei He, NWEA

One of the complex issues surrounding accommodations research is controlling for potential confounders. Here, we used propensity score matching to create groups of accommodations-eligible examinees who are equivalent on all measured individual, school, and district characteristics, except their use of accommodations during testing. Their reading performances were compared.

Comprehensive Partitioning of Student Achievement Variance to Inform Equitable Policy Design
Kyle Nickodem, University of Minnesota; Michael Rodriguez, University of Minnesota

So that large-scale assessments positively impact teaching and learning, we must comprehensively partition variance and evaluate assessment-accountability policy targets. We find about 20% of variance in a state assessment is between schools (80% within), of which 70% is explained by school demographics. Practice and policy implications are explored.

Ensuring the comparability of modified tests administered to special populations
Phoebe Winter, Independent Consultant; Mark Hansen, University of California, Los Angeles; Michelle McCoy, University of California, Los Angeles

Extensive modifications are often necessary to make a test forms accessible to certain populations. Such changes, however, can undermine comparability of test results. An approach for obtaining comparable classifications from modified forms is proposed and applied to a test of English Language Proficiency administered to blind and low vision students.

Monday, April 16, 2018
4:05-6:05pm, Majestic I, Coordinated Sessions, M6

Challenges, Issues and Opportunities in Using Response Process Data in Improving Measurement

Session Chair: Kadriye Ercikan, Educational Testing Service and UBC
Panelist: Kadriye Ercikan, Educational Testing Service and UBC
Panelist: James Pellegrino, University of Illinois Chicago
Panelist: Roy Levy, Arizona State University
Panelist: Michelle La Mar, Educational Testing Service
Panelist: Robert Mislevy, Educational Testing Service

This session will be a panel discussion on use of response process data in improving measurement. Key issues in using response process data were highlighted in the 2017 NCME volume edited by Ercikan & Pellegrino. The presenters identified below will speak to the five issues identified in the first half of the session. The audience will have a chance to send comments and questions to the chair during the presentation via twitter/email or through the NCME conference app. Panelists will respond to the audience questions in the second half of the session.

Monday, April 16, 2018
4:05-6:05pm, Plymouth, Individual Presentations, M7

Applications: Understanding Examinee Performance

Session Discussant: James Olson, Renaissance Learning

Typical Errors in English Summarizing Test Items for L2 Learners

Takahiro Terao, Nagoya University; Hidetoki Ishii, Nagoya University

This study aimed to examine typical errors made by L2 learners in English reading and summarizing test items. While less proficient test takers tended to include redundant or overly detailed information in their summaries, proficient test takers wrote summaries that included a different perspective from that of the author.

Analyzing Speech Features on Varied Item Types and First Languages

Lei Chen, Educational Testing Service; Guangming Ling, Educational Testing Service

To better understand the automated and human scoring of speaking responses, we analyzed the relationships between machine generated features, item types, and speaker's first language. We found feature values differed between item types and among first language groups, after controlling for human scores.

Diagnostic Assessment with Learning Tools to Improve the 3D Spatial Rotation Skills

Shiyu Wang, University of Georgia; Martha Carr, University of Georgia; Qi Wang, University of Florida

A computer-based 3-D spatial skills cognitive diagnostic assessment with learning tools is developed and evaluated through US and Chinese undergraduate students in this study. The proposed diagnostic assessment can help students and teachers improve spatial skills more efficiently within the context of a microgenetic study.

Using Automated Feedback to Support Students' Written Scientific Argumentation

Mengxiao Zhu, Educational Testing Service; Hee-Sun Lee, The Concord Consortium; Ting Wang, Educational Testing Service; Ou Liu, Educational Testing Service; Vinetha Belur, Educational Testing Service; Amy Pallant, The Concord Consortium

This study investigates how automated scoring and feedback can support students' construction of written scientific arguments. Using the log data that recorded argumentation scores as well as argument submission and revision records, we examined students' responses to the feedback and how their revisions related to their argumentation scores.

Measuring Science Proficiency through a More Authentic Virtual Science Laboratory

Shu-Kang Chen, Educational Testing Service; Lei Liu, Educational Testing Service; Timothy Fiser, Educational Testing Service; Katherine Castellano, Educational Testing Service; Raymond De Hont, WestEd; Delano Hebert, Educational Testing Service; Kenneth Llort, Educational Testing Service

The Virtual Science Laboratory is an open-ended, authentic 3D laboratory prototype with interactive supplies and scientifically accurate simulated phenomena to examine students' conceptual understanding and doing science. Preliminary outcomes indicated that students had limited understanding to identify chemical changes from evidence and demonstrating a range of reasoning skills.

Monday, April 16, 2018
4:05-6:05pm, Manhattan, Individual Presentations, M8

Modeling, Mediating, and Explaining DIF

Session Discussant: Bruno Zumbo, University of British Columbia

A Regularization Procedure to Detect DIF using Generalized Linear Mixed Models
Jing Jiang, Boston College; Zhushan Li, Boston College

This paper uses generalized linear mixed models to model DIF without the assumption that all items except the studied item should be invariant over groups, since all DIF parameter can be estimated simultaneously. Also, a regularization approach is introduced to solve the estimation problem and to identify the DIF items.

A Graphical Simulator for Exploring Differential Ityem Functioning
Qing Xie, University of Iowa; Terry Ackerman, ACT

The Graphical DIF Simulator (GDS) allows practitioners to explore how DIF can occur when fitting a unidimensional IRT model to two-dimensional data. Users can manipulate the underlying distributional characteristics of Reference and Focal groups, and parameters of a suspect item and observe the resulting ICCs before and after rescaling.

Comparison of MIMIC Model and HGLM to Detect and Mediate DIF
Kevin Krost, Virginia Tech

This study sought to detect and mediate gender-based differential item functioning among mathematics using the multiple indicators, multiple causes model and the hierarchical generalized linear model. Several items exhibited DIF, however the effect of gender was also mediated in several items by mathematics attitudinal scales, indicating spuriousness.

Gender Invariance on the Test-specific Student Opinion Scale
Derek Sauder, James Madison University

The test-specific version of the Student Opinion Scale (SOS) requires psychometric study before widespread use. For example, males and females may differ in how they interpret effort and test importance for various test subjects. Thus, the measurement invariance of the test-specific SOS was examined via confirmatory factor analysis.

Monday, April 16, 2018
4:05-6:05pm, Ambassador II, Coordinated Sessions, M9

Maintaining quality assessments in the face of change

Session Chair: Thanos Patelis, Human Resources Research Organization
Session Discussant: Mark Raymond, National Board of Medical Examiners

Criteria and guidelines for evaluating the quality of assessments have been provided and are currently used in many contexts. However, maintaining an assessment's quality is an on-going challenge, especially when a test is used in a high-stakes context (e.g., to screen people for jobs or select candidates for further education). When changes to an assessment are introduced to (a) represent the evolving constructs being measured, (b) influence the learning experience, or (c) take advantage of technological advancements (i.e., computer based testing, computer adaptive testing, multi-stage testing), additional efforts are needed to evaluate its quality. The purposes of this symposium are to (a) offer suggested criteria for maintaining assessment quality over time, and (b) describe efforts that illustrate how to maintain the quality of an assessment program in the face of change. This session will share illustrative solutions to ensure changes to an assessment program maintain its quality and provide criteria that participants can consider in addressing change in their assessment programs.

Criteria for Maintaining Quality Assessments
Thanos Patelis, Human Resources Research Organization

The Redesign of the Medical College Admission Test (MCAT®)
Marc Kroopnick, Association of American Medical Colleges; Ying Jin, Association of American Medical Colleges

Monitoring the Assessment's Lifecycle: Item Development through Test Administration
Michael Hughes, Human Resources Research Organization; Bethany Bynum, Human Resources Research Organization; Sean Baldwin, Human Resources Research Organization; Marc Kroopnick, Association of American Medical Colleges; Ying Jin, Association of American Medical Colleges

Exploring the Use of a Multi-State Test
Lauress Wise, Human Resources Research Organization; Matt Swain, Human Resources Research Organization; Marc Kroopnick, Association of American Medical Colleges

Exploring the Validity of the New MCAT® Exam
Kun Yuan, Association of American Medical Colleges; Cynthia Searcy, Association of American Medical Colleges

Participant Index

A

B

Participant Index

C

Participant Index

D

Participant Index

E

F

G

Participant Index

H

Participant Index

I

J

K

Participant Index

L

Participant Index

M

Participant Index

N

O

P

Participant Index

Q

R

S

Participant Index

T

Participant Index

U

V

W

Participant Index

X

Y

Z

A | Contact Information for Individual and Coordinated Sessions First Authors

A

Ackerman, Terry
ACT
terry.ackerman@act.org

Akbay, Lokman
Mehmet Akif Ersoy University / Turkey
lokmanakbay@gmail.com

Albano, Anthony
University of Nebraska-Lincoln
albano@unl.edu

Alegre, Jan
Educational Testing Service
609/683-2830

Allalouf, Avi
National InstituteTesting & Evaluation
609-734-1389

Allen, Jeff
ACT, Inc.
jeff.allen@act.org

Allen, Jeffrey
ACT, Inc.
319.337.1657

Almonte, Debby
Educational Testing Service
609/734-1137

Anderson, Daniel
University of Oregon
daniela@uoregon.edu

Anderson, Ross
Education Policy Improvement Center
5412140949

Andrews, Benjamin
ACT
319-341-2569

Andrews, Jessica
Educational Testing Service
9192808616

Andrich, David
The University of Western Australia
david.andrich@uwa.edu.au

Arce, Alvaro
Pearson
NA

Arce-Ferrer, Alvaro
Pearson
2108070983

Arieli-Attali, Meirav
ACT
meirav.attali@gmail.com

Arneson, Amy
University of California, Berkeley
510-642-0709

Attali, Yigal
Educational Testing Service
yattali@ets.org

Attali, Yigal
Educational Testing Service
609-734-1747

Austin, Bruce
Washington State University
bwaustin@wsu.edu

B

Bachman, Lyle
University of California, Los Angeles
610-850-4778

Bailey, Alison
University of California, Los Anglese
310-825-1731

Bao, Yu
University of Georgia
yubao02@uga.edu

Bashkov, Bozhidar
American Board of Internal Medicine
bbashkov@abim.org

Bause, Inga
Leibniz-Institut für Wissensmedien, Tübingen, Germany
+49 7071 979-237

Becker, Betsy
Florida State University
bbecker@fsu.edu

Contact Information for Individual and Coordinated Sessions First Authors | B

Becker, Kirk
Pearson
kirk.becker@pearson.com

Beigman Klebanov, Beata
Educational Testing Service
609-734-1330

Bejar, Isaac
Educational Testing Service
ibejar@ets.org

Bejar, Isaac
Educational Testing Service
609-734-5196

Bell, Courtney
Educational Testing Service
609-273-6328

Belov, Dmitry
Law School Admission Council
DBelov@LSAC.org

Bertling, Maria
Harvard University
NA

Betebenner, Damian
Center for Assessment
603-516-7900

Betts, Joe
Pearson VUE
312.291.5942

Black, Beth
OfQual
02476716859

Bo, Yuanchao Emily
NWEA
emily.bo@nwea.org

Bolender, Brad
ACT
brad.bolender@act.org

Bolsinova, Maria
University of Amsterdam
+31205256584

Bolt, Daniel
University of Wisconsin, Madison
dmbolt@wisc.edu

Börkan, Bengü
Boğaziçi University
bengu.borkan@boun.edu.tr

Boyer, Michelle
University of Massachusetts and Data Recognition Corporation
mlboyer@umass.edu

Braun, Henry
Boston College
braunh@bc.edu

Brennan, Robert
University of Iowa
unknown

Brenneman, Meghan
The Enrollment Management Assocation
6093604039

Briggs, Derek
University of Colorado
303-492-6320

Brodersen, Alex
University of Notre Dame
abroders@nd.edu

Brossman, Bradley
American Board of Internal Medicine
215 399 4249

Brussow, Jennifer
University of Kansas
jbrussow@gmail.com

Bu, WenJuan
Beijing Normal University Collaborative Innovation Center of Assessment toward Basic Education Quality
psybwj@163.com

Buchholz, Janine
German Institute for International Educational Research (DIPF)
buchholz@dipf.de

Buckendahl, Chad
ACS Ventures, LLC
402.770.0085

Bukhari, Nurliyana
Universiti Utara Malaysia
nurliyanabukhari@gmail.com

C | Contact Information for Individual and Coordinated Sessions First Authors

Bulus, Metin
University of Missouri
mbnt9@mail.missouri.edu

Bunch, Michael
Measurement Incorporated
unknown

Burkhardt, Amy
University of Colorado, Boulder
amy.burkhardt@colorado.edu

Burstein, Jill
Educational Testing Service
609-734-5823

Buzick, Heather
Educational Testing Service
hbuzick@ets.org

C

Cahill, Aoife
Educational Testing Service
(609)-734-1356

Cai, Li
University of California, Los Angeles
markhansen@ucla.edu

Cai, Liuhan
University of Nebraska-Lincoln
cliuhan@gmail.com

Cancado, Luciana
University of Wisconsin-Milwaukee
cancado@uwm.edu

Cao, Yi
Educational Testing Service
ycao@ets.org

Cardwell, Ramsey
University of North Carolina Greensboro
336-521-2263

Casabianca-Marshall, Jodi
Educational Testing Service
609-524-8134

Castellano, Katherine
Educational Testing Service
415-645-8449

Cetin-Berber, Dee Duygu
University of Florida
dcetinberber@ufl.edu

Chang, Meng-I
Southern Illinois Uiniversity Carbondale
mengi@siu.edu

Chao, Hsiu-Yi
National Chung Cheng University
hsiuyi1118@gmail.com

Chattergoon, Rajendra
University of Colorado Boulder
rajendra.chattergoon@colorado.edu

Chatterji, Madhabi
Columbia University, Teachers College
mb1434@tc.columbia.edu

Chen, Dandan
University of Delaware
chendan@udel.edu

Chen, Haiqin
American Dental Association
chen.haiqin2010@gmail.com

Chen, Jing
National Center for Education Statistics
202-245-8324

Chen, Jyun-Hong
National Sun Yat-sen University
horishana@gmail.com

Chen, Lei
Educational Testing Service
LChen@ets.org

Chen, Lida
The University of Iowa
lida-chen@uiowa.edu

Chen, Pei-Hua
National Chiao Tung University
peihuamail@gmail.com

Chen, Shu-Kang
Educational Testing Service
schen@ets.org

Cheng, Yi-Ling
Michigan State University
chengyil@msu.edu

Chia, Magda
Stanford University
111-11-1111

Chien, Yuehmei
Pearson
yuehmei.chien@pearson.com

Chiu, Pei-Hsuan
University of Connecticut
pei-hsuan.chiu@uconn.edu

Cho, Youngwoo
ACT
319-341-2407

Choe, Edison
Graduate Management Admission Council
echoe@gmac.com

Choi, Hye-Jeong
University of Georgia
hjchoi1@uga.edu

Choi, Ikkyu
Educational Testing Service
609-734-5163

Choi, Jinah
The University of Iowa
jinah-choi@uiowa.edu

Choi, Jinnie
Pearson
jinnie.choi@gmail.com

Choi, Seung
ACT, Inc.
(831) 383-5041

Chu, Man-Wai
University of Calgary
1-403-220-2579

Circi, Ruhan
American Institutes for Research
rcirci@air.org

Clark, Amy
University of Kansas
akclark@ku.edu

Clauser, Amanda
National Board of Medical Examiners
215-495-1477

Clauser, Jerome
American Board of Internal Medicine
jclauser@abim.org

Clifford, Ian
Prometric
ian.clifford@prometric.com

Cohen, Allan
University of Georgia
(609)-734-1356

Cohen, Yoav
NITE
972-2-6759555

Coiro, Julie
University of Rhode Island
401-874-4872

Cole, Ki
Oklahoma State University
ki.cole@okstate.edu

Colvin, Kimberly
University at Albany, SUNY
kcolvin@albany.edu

Correnti, Richard
University of Pittsburgh
412-400-2656

Croft, Michelle
ACT, Inc.
michelle.croft@act.org

Cui, Ying
Unviersity of Alberta
1-780-492-5245

Cukadar, Ismail
Florida State University
ic14d@my.fsu.edu

Curtis, Nicholas
James Madison University
curtisna@jmu.edu

D

Dadey, Nathan
The National Center for the Improvement of Educational Assessment, Inc.
ndadey@nciea.org

E | Contact Information for Individual and Coordinated Sessions First Authors

Dai, Shenghai
Washington State Unviersity
s.dai@wsu.edu

Damböck, Barbara
Akademie Dillingen, Germany
610-850-1168

Davenport, Ernest
University of Minnesota
612-624-1040

Davis, James
University of North Carolina at Greensboro
jrdavis6@uncg.edu

Davis, Jennifer
NABP
8473914521

Davis, Laurie
ACT, Inc.
laurie.davis@act.org

Davis-Becker, Susan
ACS Ventures, LLC
sdavisbecker@acsventures.com

Davison, Mark
University of Minnesota
612-624-1327

De Boeck, Paul
Ohio State University
(614) 292-4131

Deane, Paul
Educational Testing Service
609/734-1927

DeCarlo, Lawrence
Teachers College, Columbia University
decarlo@tc.edu

Denbleyker, Johnny
Houghton Mifflin Harcourt
lakeway01@yahoo.com

Deng, Sien
University of Wisconsin-Madison
sdeng7@wisc.edu

Deng, Weiilng
Educational Testing Service
WDeng@ets.org

Diao, Hongyu
University of Massachusetts Amherst
555-555-5555

Dimitrov, Dimiter
National Center for Assessment in Saudi Arabia
ddimitro@gmu.edu

Dong, Dongsheng
University of Washington
dongsd@uw.edu

Donoghue, John
Educational Testing Service
jdonoghue@ets.org

Draney, Karen
University of California, Berkeley
(510) 642-7968

Dunn, Jennifer
Questar Assessment, Inc.
603-516-7900

Dunn, Karen
British Council
karen.dunn@britishcouncil.org

E

Eichmann, Beate
German Institute for International Educational Research
beate.eichmann@dipf.de

Eklöf, Hanna
Umeå University
46 90 786 50 00

Embretson, Susan
Georgia Institute of Technology
4043850501

Embretson, Susan
Georgia Institute of Technology
4135450564

F

Fabrey, Larry
PSI Services
913-895-4706

Fan, Yuyu
Fordham University
yuyufan3@gmail.com

Farley, Daniel
University of Oregon
dfarley@uoregon.edu

Fechter, Tia
Pacific Metrics
tiacorliss@hotmail.com

Feinberg, Richard
National Board of Medical Examiners
rfeinberg@nbme.org

Feng, Gary
Educational Testing Service
609/734-1928

Feng, Jessica
Educational Testing Service
jfeng3@wellesley.edu

Feng, Luxi
Texas A&M University-College Station
sarah.feng.89@gmail.com

Feng, Yanan
Indiana University Bloomington
feng8@indiana.edu

Ferrara, Steve
Measured Progress
410-707-8059

Fina, Anthony
Iowa Testing Programs
anthony-fina@uiowa.edu

Finn, Bridgid
Educational Testing Service
609-252-8324

Fischer, Luise
University of Bamberg
fischer.luise@gmail.com

Fiser, Timothy
Educational Testing Services
tfiser@ets.org

Flanagan, Kathleen
Massachusetts Department of Elementary-and-Secondary-Education
781-338-3625

Fleckenstein, Johanna
University of Kiel
0049-4318801309

Forzani, Francesca
Teaching Works
7346472446

Fu, Zhihui
Northeast Normal University , Shenyang Normal University
fuzhihui2001@163.com

Fujimoto, Ken
Loyola University Chicago
kfujimoto@luc.edu

Fukuhara, Hirotaka
Pearson
Hiro.Fukuhara@Pearson.com

Furter, Robert
The American Board of Pediatrics
rfurter@abpeds.org

G

Gambrell, James
ACT, Inc.
James.Gambrell@act.org

Gao, Furong
Pacific Metrics Corporation
kyoko.ito.civ@mail.mil

Gao, Jie
Educational Testing Service
609/734-1815

Gao, Lingyun
Measured Progress, Inc.
lingyun_gao@hotmail.com

Gerasimova, Daria
George Mason University
dgerasim@masonlive.gmu.edu

Ghosh, Sattik
UC Davis
stkghosh@ucdavis.edu

Gialluca, Kathleen
Pearson VUE
952.681.3856

H | Contact Information for Individual and Coordinated Sessions First Authors

Gillespie, Sally
Educational Testing Service
609/683-2407

Glazer, Nancy
Educational Testing Service
6097345413

Gochyyev, Perman
University of California, Berkeley
perman@berkeley.edu

Godoy, María Inés
Pontificia Universidad Católica de Chile/MIDE UC
migodoy1@uc.cl

Goldstein, Harold
Baruch College - CUNY
(646) 312-3820

Gong, Brian
Center for Assessment
bgong@nciea.org

Goodman, Joshua
NCCPA
3367400636

Grady, Matthew
American Dental Association
gradym@ada.org

Graesser, Arthur
University of Memphis
901-678-4857

Graf, Edith
Educational Testing Service
agraf@ets.org

Greiff, Samuel
University of Luxembourg
+352-466644-9245

Grochowalski, Joseph
The College Board
joe.grochowalski@gmail.com

Guan, Li
University of Georgia
aguan0215@gmail.com

Guerere, Claudia
Schroeder Measurement Technologies
cguerere@mail.usf.edu

Guo, Hongwen
Educational Testing Service
hguo@ets.org

Guo, Qi
Unviersity of Alberta
1-780-492-5245

Guo, Shaoyang
Jiangxi Normal University
syguo1992@outlook.com

Guzman-Orth, Danielle
Educational Testing Service
415-645-8457

H

Haag, Nicole
Institute for Educational Quality Improvement
nicole.haag@iqb.hu-berlin.de

Haisfield, Lisa
ACT
Liza_Haisfield@hotmail.com

Hambleton, Ronald
University of Massachusetts, Amherst
unknown

Hamilton, Laura
RAND Education
(310) 451-6913

Hansen, Eric
Educational Testing Service
6097345413

Hansen, John
Harvard University
253.365.2423

Hansen, Mark
University of California, Los Angeles
markhansen@ucla.edu

Hauenstein, Clifford
Georgia Institute of Technology
4043850501

Hauger, Jeffrey
New Jersey Department of Education
877-900-6960

He, Qiwei
Educational Testing Service
+1-6092436542

He, Yi
ACT
uiheyi@gmail.com

Herborn, Katharina
University of Luxembourg
+352 46 66 44 5578

Himelfarb, Igor
The National Board of Chiropractic Examiners
ihimelfarb@nbce.org

Ho, Andrew
Harvard University
617-496-2408

Ho, Margaret
CRESST
(323) 657-3096

Hochstetter, Angela
Minnesota Department of Education
angela.hochstetter@state.mn.us

Hu, Bo
University of Kansas
who.bo@ku.edu

Huff, Kristen
Curriculum Associates
7184502205

Huggins-Manley, Anne Corinne
University of Florida
ahuggins@coe.ufl.edu

Hughes, Michael
Human Resources Research Organization
703-706-5663

Huh, Nooree
ACT, Inc.
nooree.huh@act.org

Hummel, Steven
American Institutes for Research
202-403-6420

Huo, Yan
Educational Testing Service
yhuo@ets.org

Hurtz, Gregory
PSI Services LLC
ghurtz@psionline.com

Hurwitz, Michael
College Board
617.495.1005

I

Iaconangelo, Charles
Pharmerit International
charles.iaconangelo@gmail.com

Insko, William
Houghton Mifflin Harcourt
bill.insko@hmhco.com

Insko, Jr., William
Houghton Mifflin Harcourt
630-467-6163

J

Janssen, Rianne
KU Leuven (Belgium)
rianne.janssen@kuleuven.be

Jewsbury, Paul
Educational Testing Service
pjewsbury@ets.org

Jia, Yue
Educational Testing Service
609-734-1224

Jiang, Bingnan
ACT. Inc.
(303) 607-3441

Jiang, Jing
Boston College
jiangjc@bc.edu

Jiang, Tao
American Institutes for Research
xd.turtle@gmail.com

Jiang, Yanlin
Educational Testing Service
yjiang@ets.org

K | Contact Information for Individual and Coordinated Sessions First Authors

Jiang, Yanming
Educational Testing Service
yxjiang@ets.org

Jiang, Zhehan
University Of Kansas
zjiang4@ku.edu

Jiao, Hong
University of Maryland, College Park
hjiao@umd.edu

Jing, Shumin
University of Iowa
shumin-jing@uiowa.edu

Johnson, Matthew
Teachers College, Columbia University
212-678-3429

Jones, Andrew
American Board of Surgery
ajones@absurgery.org

Jones, Eli
University of Missouri
eliandrewjones@gmail.com

Jorion, Natalie
PearsonVUE
talie.jorion@gmail.com

Ju, Unhee
Michigan State University
juunhee@msu.edu

Jung, HyunJoo
University of Massachusetts Amherst
hyunjoo.jung2@gmail.com

K

Kaira, Leah
ES Pearson
4132567239

Kaliski, Pamela
College Board
pamela.kaliski@gmail.com

Kane, Michael
Educational Testing Service
unknown

Kang, Youngsoon
University of Minnesota
612-626-1662

Kao, Shu-chuan
Pearson
kaoshuch@msn.com

Kapoor, Shalini
ACT
319-337-1946

Kara, Yusuf
Southern Methodist University
ykara88@gmail.com

Karatoprak Ersen, Rabia
The University of Iowa
karatoprak.rabia@gmail.com

Keller, Stefan
Fachhochschule Nordwestschweiz
0041-616901914

Kennet-Cohen, Tamar
National Institute for Testing and Evaluation, Israel
tami@nite.org.il

Kenyon, Dorry
Center for Applied Linguistics
202-362-0700

Ketterlin Geller, Leanne
Southern Methodist University
lkgeller@smu.edu

Keum, EunHee
UCLA/CRESST
keum@cresst.org

Khan, Saad
Educational Testing Service
+ 1 6092528427

Kim, Ahyoung Alicia
University of Wisconsin-Madison
608-890-1379

Kim, Do-Hong
Augusta University
kimdohong@gmail.com

Kim, Doyoung
NCSBN
312.291.5942

Kim, Han Yi
Measured Progress
hanyi.kim.ui@gmail.com

Kim, Hyung Jin
The University of Iowa
hyungjin-kim@uiowa.edu

Kim, Kyung Yong
University of North Carolina at Greensboro
k.kyungyong@gmail.com

Kim, Meereem
University of Georgia
always0531@gmail.com

Kim, Seock-Ho
The University of Georgia
shkim@uga.edu

Kim, Sewon
Michigan State University
kimsewon@msu.edu

Kim, Stella
The University of Iowa
stella-kim@uiowa.edu

Kim, Young Yee
American Institues for Research
ykim@air.org

Kim, YoungKoung
The College Board
ykkim08@gmail.com

Klieger, David
Educational Testing Service
dklieger@ets.org

Ko, Inah
University of Michigan
inahko@umich.edu

Kolstad, Andrew
P20 Strategies LLC
ajk95@columbia.edu

Kosh, Audra
MetaMetrics, Inc.
audrakosh@gmail.com

Kroehne, Ulf
DIPF (German Institute for International Educational Research)
kroehne@dipf.de

Kroopnick, Marc
Association of American Medical Colleges
202-828-0968

Krost, Kevin
Virginia Tech
kevinkrost@vt.edu

Krueger, Maleika
Fachhochschule Nordwestschweiz
0041-616901914

Kuhfeld, Megan
NWEA
megan.kuhfeld@gmail.com

Kuhn, Christiane
Johannes Gutenberg-University Mainz (Germany), Department of Business and Economics Education
info@elmawi.de

Kuo, Tzu-Chun
American Institute for Research
tkuo@air.org

Kyllonen, Patrick
Educational Testing Service
609-734-1056

L

Laitusis, Cara
Educational Testing Service
6097345413

Lakin, Joni
Auburn University
111-11-1111

Lamsal, Sunil
Pearson VUE
lamsals@hotmail.com

Lane, Suzanne
University of Pittsburgh
412-648-7095

Lathrop, Quinn
Pearson Advanced Computing and Data Science Lab
quinn.lathrop@gmail.com

Lau, Clarissa
University of Toronto
clarissa.lau@mail.utoronto.ca

L | Contact Information for Individual and Coordinated Sessions First Authors

Lee, HyeSun
California State University Channel Islands
hyesun.lee@csuci.edu

Lee, Moonsoo
Korea Institute for Curriculum and Evaluation(KICE)
mslee9@kice.re.kr

Lee, Won-Chan
The University of Iowa
319-335-5546

Lee, Yi-Hsuan
Educational Testing Service
ylee@ets.org

Lee, Yi-Hsuan
Educational Testing Service
609/734-1176

Lei, Ming
American Institutes of Research
mlei@air.org

Leighton, Jacqueline
University of Alberta
1-780-492-5245

Lewis, Daniel
ACT, Inc.
831-383-5043

Li, Anqi
University of Illinois at Urbana-Champaign
anqili4@illinois.edu

Li, Caihong
University of Kentucky
caihong.li@uky.edu

Li, Chen
University of Maryland, College Park
lichen1210@gmail.com

Li, Feiming
Zhejiang university of Technology
404688102@qq.com

Li, Hongli
Georgia State University
hli24@gsu.edu

Li, Jie
ACT, Inc.
lijdbc@gmail.com

Li, Tianli
ACT Inc.
tianli.li@act.org

Li, Tongyun
Educational Testing Service
tli002@ets.org

Li, Xiao
University of Illinois at Urbana-Champaign
xiaoli20@illinois.edu

Li, Xin
ACT, Inc.
xin.li@act.org

Li, Yanmei
Educational Testing Service
yxli@ets.org

Li, Zhen
eMetric
zli@emetric.net

Li, Zhen
Government of Newfoundland and Labrador
liza0616@hotmail.com

Li, Zhushan
Boston College
zhushan.li@bc.edu

Liang, Longjuan
Educational Testing Service
609-734-5220

Liao, Dandan
University of Maryland, College Park
dandanl@umd.edu

Liao, Manqian
University of Maryland College Park
mancyliao@gmail.com

Liaw, Yuan-Ling
University of Oslo Centre for Educational Measurement
y.l.liaw@cemo.uio.no

Lim, Hwanggyu
University of Massachusetts Amherst
hglim83@gmail.com

Lin, Chih-Kai (Cary)
American Institutes for Research (AIR)
clin@air.org

Lin, Meiko
Teachers College, Columbia University
ml2734@columbia.edu

Lin, Ye
University of Iowa
ye-lin@uiowa.edu

Ling, Guangming
Educational Testing Service
609-734-5594

Liu, Lei
Educational Testing Service
609-734-5183

Liu, Ou
Educational Testing Service
6097341049

Liu, Qiongqiong
National Board of Osteopathic Medical Examiners
echo_910@hotmail.com

Liu, Xin
Ascend Learning
Lucy.Xin.Liu@gmail.com

Liu, Yaping
Beijing Normal University
18813005963@163.com

Livingston, Samuel
Educational Testing Service
slivingston@ets.org

Llosa, Lorena
New York University
2129985485

Lochbaum, Karen
Pearson
720-476-3517

Lopez, Alexis
Educational Testing Service
111-11-1111

Lottridge, Sue
ACT
(720)-544-6187

Lu, Ru
Educational Testing Service
rlu@ets.org

Lu, Yang
ACT
319-341-2915

Luciw-Dubas, Ulana
National Board of Medical Examiners
udubas@nbme.org

Luecht, Richard
University of North Carolina Greensboro
336-404-0746

Luo, Xiao
National Council of State Boards of Nursing
xluo1986@gmail.com

Luo, Xin
Uber
charonluo@gmail.com

Luo, Yong
National Center for Assessment, Saudi Arabia
jackyluoyong@gmail.com

M

Ma, Wenchao
The University of Alabama
wenchao.ma@ua.edu

Ma, Ye
the University of Iowa
ye-ma@uiowa.edu

MacArthur, Charles
University of Delaware
302-831-4572

Maddox, Bryan
University of East Anglia
+44 (0)1603 59 3380

Madison, Matthew
University of California, Los Angeles
mjmadison@ucla.edu

Maeda, Hotaka
University of Wisconsin-Milwaukee
hotaka.maeda@gmail.com

Malatesta, Jaime
The University of Iowa
jaime-malatesta@uiowa.edu

M | Contact Information for Individual and Coordinated Sessions First Authors

Man, Kaiwen
University of Maryland College Park
kman@umd.edu

Marcoulides, Katerina
University of Florida
kmarcoul@asu.edu

Maris, Gunter
ACTNext
319.341.2449

Marksteiner, Tamara
University of Mannheim
+49 621-181 2210

Martineau, Joseph
Center for Assessment
517-410-5220

Martinez, Jose Felipe
University of California, Los Angeles
310-794-1853

Mason, James
University of California, Berkeley
510-642-0709

Mattern, Krista
ACT, Inc.
319.337.1182

Mazzeo, John
Educational Testing Service
609-921-9000

Mbekeani, Preeya
Harvard University
617.495.1005

McCaffrey, Daniel
Educational Testing Service
609-252-8404

McCoy, Michelle
CRESST
310-625-8635

McGlone, Moni
University of Wisconsin-Madison
111-11-1111

McMurran, Meaghan
University of California Riverside
mmcmu001@ucr.edu

McNaughton, Tara
Measurement Incorporated
tmcnaughton@measinc.com

Mee, Janet
National Board of Medical Examiners
jmee@nbme.org

Meijer, Rob
University of Groningen, The Netherlands
+31 50 36 36339

Meisner, Richard
ACT, Inc.
meisner.rick@gmail.com

Meng, Huijuan
GMAC
7036689749

Meyer, Patrick
University of Virginia
jpm4qs@virginia.edu

Michaelides, Michalis
University of Cyprus
michalim@ucy.ac.cy

Milligan, Sandra
University of Melbourne
s.milligan@unimelb.edu.au

Minchen, Nathan
Rutgers, The State University of New Jersey
nathanminchen@gmail.com

Molenaar, Dylan
University of Amsterdam
+31205256584

Monroe, Scott
UMass Amherst
smonroe@educ.umass.edu

Morell, Monica
University of Maryland
mmorell@umd.edu

Morris, Scott
Illinois Institute of Technology
scott.morris@iit.edu

Morrison, Kristin
ACT, Inc.
Kristin.Morrison@act.org

Moses, Tim
The College Board
215 801-0476

Moxley, Joe
University of South Florida
813.404.9734

Muckle, Timothy
National Board of Certification and Recertification for Nurse Anesthetists
tmuckle@nbcrna.com

Muntean, William
Pearson Vue
william.muntean@pearson.com

Myers, Aaron
James Madison University
myers2aj@jmu.edu

N

Nash, Brooke
University of Kansas
bnash@ku.edu

Naumann, Alexander
German Institute for International Educational Research (DIPF)
naumanna@dipf.de

Nickodem, Kyle
University of Minnesota
nicko013@umn.edu

Niessen, Susan
University of Groningen
a.s.m.niessen@rug.nl

Niu, Luping
The University of Texas at Austin
NEWL787@gmail.com

Norris, Mary
Virginia Tech
mnorris@vt.edu

O

Oh, Hyeonjoo
Educational Testing Service
hoh@ets.org

Olgar, Suleyman
Florida Department of Education
3216528131

Oliveri, Maria Elena
Educational Testing Service
moliveri@ets.org

O'Neill, Thomas
American Board of Family Medicine
555-555-5555

Ortiz, Samuel
St. John's University
718-990-5388

P

Padellaro, Frank
University of Massachusetts Amherst
fpadellaro@umass.edu

Paek, Pamela
ACT, Inc.
Pamela.Paek@act.org

Pak, Seohong
National Board of Medical Examiners
seoccaatt@gmail.com

Pan, Qianqian
Achievement and Assessment Institute, University of Kansas
qpan@ku.edu

Papanastasiou, Elena
University of Nicosia
22842316

Park, Jung Yeon
University of Leuven
ellie.park@kuleuven.be

Park, Jungkyu
McGill University
jungkyu.park@mail.mcgill.ca

Q | Contact Information for Individual and Coordinated Sessions First Authors

Park, Seohee
University of Iowa
seohee-park@uiowa.edu

Park, Yoon Soo
University of Illinois at Chicago
yspark2@uic.edu

Patelis, Thanos
Human Resources Research Organization
518-545-8253

Patton, Elizabeth
University of North Carolina Greensboro
eapatton@uncg.edu

Pecheone, Raymond
Stanford University
650-892-5956

Peng, Fang
University of Illinois at Chicago
pfrenee@gmail.com

Peters, Stephanie
Educational Testing Service
speters@ets.org

Petway, Kevin
The Enrollment Management Assocation
7033177075

Pham, Duy
University of Massachusetts Amherst
dpham@umass.edu

Poggio, John
University of Kansas
jpoggio@ku.edu

Polyak, Steve
ACTNext
319.341.2449

Powers, Sonya
ACT
520-269-5412

Pöysä-Tarhonen, Johanna
University of Jyväskylä
+358 400 248 124

Priniski, Stacy
University of Wisconsin, Madison
608-444-6352

Ptukhin, Yevgeniy
Southern Illinois University Carbondale
ptukyevg@siu.edu

Puhan, Gautam
Educational Testing Service
609-734-5240

Q

Qian, Hong
NCSBN
3125253721

Qian, Jiahe
Educational Testing Service
jqian@ets.org

Qiao, Xin
University of Maryland College Park
xin.qiao56@gmail.com

Qiu, Xue-Lan
Department of Psychology, The Education University of Hong Kong, Hong Kong
psyxlq@gmail.com

Quesen, Sarah
Pearson
sarah.quesen@gmail.com

Qureshi, Farah
Educational Testing Service
609/734-1170

R

Raborn, Anthony
University of Florida
lordmaxwell@ufl.edu

Rahman, Taslima
National Center for Educational Statistics
(202) 245-6514

Raymond, Mark
National Board of Medical Examiners
mraymond@nbme.org

Reddy, Linda
Rutgers, the State University of New Jersey
adam.lekwa@rutgers.edu

Reid, Aileen
University of North Carolina at Greensboro
amreid3@uncg.edu

Ren, Hao
ACT. Inc.
(517) 525-8506

Reyes, Lisa
Measurement Incorporated
lreyes@measinc.com

Rickels, Heather
University of Iowa, Iowa Testing Programs
heather-rickels@uiowa.edu

Rijmen, Frank
AIR
frankrijmen@hotmail.com

Rikoon, Sam
Educational Testing Service
6092528613

Riordan, Brian
Educational Testing Service
609-524-8036

Rios, Joseph
Educational Testing Service
jrios@ets.org

Roberts, James
Georgia Institute of Technology
james.roberts@psych.gatech.edu

Rodriguez, Michael
University of Minnesota
mcrdz@umn.edu

Rome, Logan
Curriculum Associates
logan_rome@yahoo.com

Rotou, Ourania
Educational Testing Service
orotou2001@yahoo.com

Runyon, Christopher
The University of Texas at Austin
runyon.christopher@utexas.edu

Rupp, Andre
Educational Testing Service
609-252-8545

Russell, Morgan
Mursion
(415) 624-3837

Rust, Keith
WESTAT
301-251-8278

Rutkowski, Leslie
University of Oslo
leslie.rutkowski@cemo.uio.no

S

Sahin, Fusun
American Institues for Research
202.403.5224

Sanchez, Edgar
ACT, Inc
edgar.sanchez@act.org

Sari, Halil
Kilis 7 Aralik University
hisari87@gmail.com

Sato, Yoshikazu
Kyushu University
ysato@artsci.kyushu-u.ac.jp

Sauder, Derek
James Madison University
sauderdc@dukes.jmu.edu

Schlax, Jasmin
Johannes Gutenberg-University Mainz
jasmin.schlax@uni-mainz.de

Schneider, Bertrand
Harvard University
NA

Schulz, Matthew
Smarter Balanced Consortia Assessments
831-646-6404

Schulze, Daniel
Freie Universität Berlin
sulzedan@hu-berlin.de

Scoular, Claire
University of Melbourne
c.scoular@unimelb.edu.au

T | Contact Information for Individual and Coordinated Sessions First Authors

Sessoms, John
Measured Progress
jcsessom@uncg.edu

Sgammato, Adrienne
Educational Testing Service
asgammato@ets.org

Shear, Benjamin
University of Colorado Boulder
benjamin.shear@colorado.edu

Shen, Yawei
The University of Georgia
ys37335@uga.edu

Shermis, Mark
University of Houston-Clear Lake
954-899-8069

Shin, Ching-Wei
Pearson
cshin0803@gmail.com

Shin, Hyo Jeong
Educational Testing Service
hshin@ets.org

Shin, Nami
University of California, Los Angeles
nami0623@gmail.com

Shin, Nami
CRESST
310-267-4476

Shojaee, Mahnaz
University of Alberta, Educational Psychology, "Measurement, Evaluation & Cognition"
mshojaee@ualberta.ca

Shu, Zhan
Educational Testing Service
zshu@ets.org

Sinharay, Sandip
Educational Testing Service
ssinharay@ets.org

Sireci, Stephen
University of Massachusetts Amherst
4135450564

Smiley, Whitney
American Board of Internal Medicine
whitknee48@gmail.com

Smith, Weldon
University of Nebraska Lincoln, Buros Center for Testing
weldon@huskers.unl.edu

Solano-Flores, Guillermo
Stanford University
(650) 723-2109

Song, Victoria
Fordham University
vsong2@fordham.edu

Srinivasan, Jayashri
University of California, Los Angeles
jsrini@ucla.edu

Stancavage, Fran
American Institutes for Research
650-400-9575

Su, Ya-Hui
National Chung Cheng University
psyyhs@ccu.edu.tw

Sun, Xiaojian
Beijing Normal University
sun.xiaojian@outlook.com

Sun, Yan
Rutgers University
yan.sun@rutgers.edu

Sussman, Joshua
University of California, Berkeley
4155317854

Suzuki, Lisa
NYU Steinhardt
212 998 5575

Svetina, Dubravka
Indiana University
dsvetina@indiana.edu

T

Tanaka, Victoria
The University of Georgia
vtanaka@uga.edu

Tao, Shuqin
Curriculum Associates
shuqin.tao@gmail.com

Terao, Takahiro
Nagoya University
terao.takahiro@a.mbox.nagoya-u.ac.jp

Terzi, Ragip
The Turkish Ministry of National Education
terziragip@gmail.com

Thompson, William
University of Kansas - Dynamic Learning Maps
wjakethompson@gmail.com

Thum, Yeow
NWEA, Portland, OR
yeow.meng@nwea.org

Tijmstra, Jesper
Tilburg University
+31 13 466 2089

Topczewski, Anna
GED Testing Service
NA

Tran, Sarah-Truclinh
NWEA
sarah.tran@nwea.org

Turner, Kyle
The University of Georgia
kturner2@uga.edu

U

Ulitzsch, Esther
Freie Universität Berlin
esther.ulitzsch@fu-berlin.de

V

van der Linden, Wim
ACT, Inc.
(831) 383-5409

Verkuilen, Jay
CUNY Graduate Center
jverkuilen@gc.cuny.edu

Vispoel, Walter
University of Iowa
walter-vispoel@uiowa.edu

von Davier, Alina
ACTNext
319.341.2449

von Davier, Matthias
NBME
mvondavier@gmail.com

W

Walker, Michael
The College Board
memwalker@gmail.com

Wan, Lei
The College Board
hyunjoo.jung2@gmail.com

Wang, Qinjun
University of Minnesota - Twin Cities
wang4314@umn.edu

Wang, Shichao
ACT, Inc
shichao.wang@act.org

Wang, Shiyu
University of Georgia
swang44@uga.edu

Wang, Ting
The American Board of Anesthesiology
ting.wang@theaba.org

Wang, Wei
Educational Testing Service
weiwang752@gmail.com

Wang, Xi
Measured Progress
smilingwx2010@gmail.com

Wang, Xiaolin
University of Kansas
xw41@indiana.edu

Wang, Xinrui
Pearson VUE
xinrui.wang2008@gmail.com

Wang, Yang
Education Analytics
cwang@edanalytics.org

W | Contact Information for Individual and Coordinated Sessions First Authors

Wang, Zhen
Educational Testing Servic
jwang68@hotmail.com

Way, Walter
College Board
240-618-5261

Weeks, Jonathan
Educational Testing Service
6099219000

Wei, Youhua Wei
Educational Testing Service
ywei@ets.org

Weiner, John
PSI Services LLC
ghurtz@psionline.com

Weir, J. B.
University of North Carolina at Greensboro
weirjb@gmail.com

Wells, Craig
University of Massachusetts Amherst
(413) 577-1726

Wendler, Cathy
Educational Testing Service
609-734-5542

Weren, Barbara
Educational Testing Service
6097345413

Westrick, Paul
ACT
paul.westrick@act.org

White, Lauren
Florida Department of Education
3014524165

White, Sheida
National Center for Education Statistics
202-245-7115

Whitehill, Jacob
Worcester Polytechnic Institute
NA

Widiatmo, Heru
ACT, Inc.
heru.widiatmo@act.org

Wihardini, Diah
Bina Nusantara University
diah.wihardini@berkeley.edu

Wikstrom, Christina
Umea University
+46 90 786 55 70

Wilson, Mark
University of California, Berkeley
510-542-4725

Winter, Phoebe
Independent Consultant
markhansen@ucla.edu

Wise, Lauress
Human Resources Research Organization
831-647-1004

Wise, Steven
NWEA
steve.wise@nwea.org

Wladis, Claire
CUNY Graduate Center
cwladis@gmail.com

Wolf, Mikyung
Educational Testing Service
mkwolf@ets.org

Wolfe, Edward
Educational Testing Service
609-524-8140

Wolkowitz, Amanda
Alpine Testing Solutions
amanda.wolkowitz@alpinetesting.com

Wong, Pamela
New York City Department of Education
pwong7@schools.nyc.gov

Woo, Ada
NCSBN
312.291.5942

Woo, Ada
NCSBN
312 525 3690

Wools, Saskia
Cito
+31 6 10963301

Wylie, Caroline
Educational Testing Service
609-510-1060

Wyse, Adam
The American Registry of Radiologic Technologists
adam.wyse@arrt.org

X

Xi, Nuo
Educational Testing Service
609-734-1895

Xia, Yan
Arizona State University
yxia@asu.edu

Xie, Aolin
Prometric. Inc
olymxie@gmail.com

Xie, Qing
U of Iowa
qing-xie@uiowa.edu

Xu, Jing-Ru
Pearson
jingruxu2013@gmail.com

Xu, Wei
National Council of State Boards of Nursing (NCSBN)
x.wei1007@gmail.com

Xu, Xueli
Educational Testing Service
609-921-9000

Y

Yang, Jing
Northeast Normal University
yangj014@nenu.edu.cn

Yee, Darrick
Harvard Graduate School of Education
darrick.yee@gmail.com

Yin, Ping
Curriculum Associates
pingyin04@hotmail.com

Yoo, Hanwook
Educational Testing Service
hyoo@ets.org

Yu, Xiaofeng
University of Notre Dame
xyu6@nd.edu

Yuan, Kun
Association of American Medical Colleges
202-828-0968

Z

Zhan, Peida
Beijing Normal University
pdzhan@gmail.com

Zhang, Caiyan
The College Board
caicaiyan.z@gmail.com

Zhang, Ci
University of Illinois at Urbana-Champaign
cizhang2@illinois.edu

Zhang, Jiahui
Michigan State University
zhang321@msu.edu

Zhang, Jiaqi
University of Cincinnati
zhangjq@mail.uc.edu

Zhang, Jin
ACT Inc.
jin.zhang@act.org

Zhang, Jinming
University of Illinois at Urbana-Champaign
jmzhang@illinois.edu

Zhang, Liru
Delaware Department of Education
302-736-3367

Zhang, Susu
University of Illinois at Urbana-Champaign
szhan105@illinois.edu

Zhang, Xinxin
University of Alberta
xinxin4@ualberta.ca

Z | Contact Information for Individual and Coordinated Sessions First Authors

Zhang, Yongmei
Beijing Academy of educational sciences
zhym72@163.com

Zhang, Zhonghua
Melbourne Graduate School of Education, University of Melbourne
chonghuachang@gmail.com

Zhao, Xinchu
University of South Carolina
xinchu@email.sc.edu

Zheng, Xiaying
University of Maryland, College Park
xyzheng86@gmail.com

Zheng, Yi
Arizona State University
yi.isabel.zheng@asu.edu

Zhou, Xuechun
NCS Pearson
xuechun.zhou@pearson.com

Zhu, Mengxiao
Educational Testing Service
mzhu@ets.org

Zijlmans, Eva
Tilburg University
e.a.o.zijlmans@tilburguniversity.edu

Zisk, Robert
Rutgers University
848-932-0642

Zlatkin-Troitschanskaia, Olga
Johannes Gutenberg-University Mainz (Germany), Department of Business and Economics Education
info@elmawi.de

Zopluoglu, Cengiz
University of Miami
305-284-5102

Zor, Selay
University of Georgia
sz37952@uga.edu

Zu, Jiyun
Educational Testing Service
jzu@ets.org

Zwick, Rebecca
Educational Testing Service
805-680-8356

NCME 2018 • Schedule-At-A-Glance

Time	Room	Type	ID	Title
Thursday, April 12				
8:00 AM–5:00 PM	Ambassador III-The Westin	TS	AA	Cognitive Diagnosis Modeling: A General Framework Approach and Its Implementation in R
8:00 AM–5:00 PM	Murray Hill East-The Hilton	TS	BB	Measuring hard-to-measure (noncognitive) skills: Social, emotional, self-management, and beyond
8:00 AM–5:00 PM	Clinton Suite-The Hilton	TS	CC	Techniques and Software for Q-Matrix Estimation and Modeling Learning in Cognitive Diagnosis
8:00 AM–5:00 PM	Murray Hill West-The Hilton	TS	DD	Using Shiny to create custom psychometric solutions
8:00 AM–12:00 PM	Madison- The Hilton	TS	EE	Computerized Multistage Adaptive Testing: Theory and Applications
8:00 AM–12:00 PM	Ambassador II-The Westin	TS	FF	Federal Education Policy as a Driver of Assessment Design (1965 to present)
8:00 AM–12:00 PM	Gibson Suite-The Hilton	TS	GG	Moving From Paper to Online Assessments: Psychometric, Content, and Classroom Considerations
1:00–5:00 PM	Gibson Suite-The Hilton	TS	HH	An Overview of Operational Psychometric Work in Real World
1:00–5:00 PM	Madison- The Hilton	TS	II	Collaborative Solution Design for Educational Measurement Challenges: Not a Spectator Sport
1:00–5:00 PM	Ambassador II-The Westin	TS	JJ	Effective Item Writing for Valid Measurement
1:00–5:00 PM	Nassau West- The Hilton	TS	KK	Practical Applications of Vertical Articulation in Standard Setting
Friday, April 13				
8:00 AM–5:00 PM	Broadway I	TS	LL	Bayesian Networks in Educational Assessment
8:00 AM–5:00 PM	Ambassador III	TS	MM	LNIRT: Joint Modeling of Responses (Accuracy) and Response Times (Speed)
8:00 AM–5:00 PM	Broadway II	TS	NN	Shadow-Test Approach to Adaptive Testing
8:00 AM–5:00 PM	Broadway III	TS	OO	Test Equating Methods and Practices
8:00 AM–12:00 PM	Belasco	TS	PP	The Stanford Education Data Archive: Using big data to study academic performance
8:00 AM–12:00 PM	Majestic I	TS	QQ	A Visual Introduction to Computerized Adaptive Testing
8:00 AM–12:00 PM	Ambassador II	TS	RR	Applying Test Score Equating Methods Using R
8:00 AM–12:00 PM	Gershwin I	TS	SS	Diagnostic Classification Models Part I: Fundamentals

CS=Coordinated Session • EB= Electronic Board Session
IS= Invited Session • PS= Paper Session • TS=Training Session

Time	Room	Type	ID	Title
1:00–5:00 PM	Belasco	TS	TT	Analyzing NAEP Data Using Plausible Values and Marginal Estimation With AM
1:00–5:00 PM	Majestic I	TS	UU	Bayesian Analysis of Response Style IRT Models Using SAS PROC MCMC
1:00–5:00 PM	Gershwin II	TS	VV	Diagnostic Classification Models Part II: Advanced Applications
1:00–5:00 PM	Majestic II	TS	WW	Evidence-Centered Design and Computational Psychometrics Solution for Game/Simulation-Based Assessments
1:00–5:00 PM	Ambassador II	TS	XX	Landing Your Dream Job for Graduate Students
4:00–7:00 PM	Minskoff			NCME Board of Directors Meeting
4:30–6:30 PM				Graduate Student Social
Saturday, April 14				
6:30 AM – 7:30 AM	Majestic I			Yoga
8:15 AM – 10:15 AM	Ambassador III	CS	A1	Are We Entering a New Era for Educational Assessment?
8:15 AM – 10:15 AM	Broadway I	CS	A2	Advances and Perspectives in Machine Scoring
8:15 AM – 10:15 AM	Broadway II	CS	A3	Understanding, Predicting, and Modifying the Performance of Human Raters
8:15 AM – 10:15 AM	Broadway III	CS	A4	What Writing Analytics Can Tell Us About Broader Success Outcomes
8:15 AM – 10:15 AM	Gershwin 2	CS	A5	Developing Simulated Performance Assessments for use in Teacher Licensure
8:15 AM – 10:15 AM	Belasco	PS	A6	Exploring Linking Designs
8:15 AM – 10:15 AM	Plymouth	PS	A7	Using TIming Data in Innovative Ways
8:15 AM – 10:15 AM	Manhattan	PS	A8	Evaluating Current and Emerging Psychometric Models and Methods
8:15 AM – 10:15 AM	Ambassador II	PS	A9	Setting Performance Standards: New Contexts and Approaches
10:35 AM – 12:05 PM	Ambassador III	IS	B1	The Past, Present, and Future of Curriculum-Based Measurement
10:35 AM – 12:05 PM	Broadway I	CS	B2	Challenges and Opportunities on International Higher Education Admission Practices
10:35 AM – 12:05 PM	Broadway II	CS	B3	Validity Considerations for New Data in Performance Learning and Assessment
10:35 AM – 12:05 PM	Broadway III	CS	B4	Experimental Design within a Survey Assessment: Learning from NAEP Digital Transition
10:35 AM – 12:05 PM	Gershwin 2	CS	B5	Digitally Simulated Science Laboratory Assessments: Differential Approaches for Analyzing Log File Data

CS=Coordinated Session • EB= Electronic Board Session
IS= Invited Session • PS= Paper Session • TS=Training Session

Time	Room	Type	ID	Title
10:35 AM – 12:05 PM	Belasco	CS	B6	Estimating Parameters in an Adaptive Context
10:35 AM – 12:05 PM	Plymouth	CS	B7	Diagnostic Classification Models: Challenges and Opportunities
10:35 AM – 12:05 PM	Manhattan	CS	B8	Validating Assessments for Particular Uses
10:35 AM – 12:05 PM	Ambassador II	PS	B9	Perspectives on Response Modeling
10:35 AM – 12:05 PM	Gershwin I	EB	B10	Electronic Board Session 1
12:25 PM – 1:55 PM	Ambassador III	CS	C1	History of Measurement from 1950 to the Present - Part 1
12:25 PM – 1:55 PM	Broadway I	CS	C2	Addressing Motivational Issues in Low-Stakes Testing: U.S. and International Perspectives
12:25 PM – 1:55 PM	Broadway II	CS	C3	Measuring Clinical Judgment in Nursing: Integrating Technology Enhanced Items
12:25 PM – 1:55 PM	Broadway III	CS	C4	Classroom Assessment and Educational Measurement
12:25 PM – 1:55 PM	Gershwin 2	IS	C5	National Association of Assessment Directors and National Council on Measurement in Education 2018—Creating the Capacity to Increase Understanding of What Works in Schools, How It's Measured and Why It Works
12:25 PM – 1:55 PM	Belasco	PS	C6	Something's Missing: Working with Incomplete Data
12:25 PM – 1:55 PM	Plymouth	PS	C7	Moving forward with MST
12:25 PM – 1:55 PM	Manhattan	PS	C8	IRT for Designing and Evaluating Tests
12:25 PM – 1:55 PM	Ambassador II	PS	C9	Reflecting on Item and Form Development
12:25 PM – 1:55 PM	Gershwin I	EB	C10	GSIC Graduate Student Poster Session 1
2:15 PM – 3:45 PM	Ambassador III	CS	D1	History of Measurement from 1950 to the Present - Part 2
2:15 PM – 3:45 PM	Broadway I	CS	D2	Using Classification-based Psychometrics in Local Assessment Systems for Feedback and Accountability
2:15 PM – 3:45 PM	Broadway II	CS	D3	Response times in educational measurement: Moving beyond the simple structure hierarchical model
2:15 PM – 3:45 PM	Broadway III	CS	D4	Measuring Collaboration and Engagement using "Big Data"
2:15 PM – 3:45 PM	Gershwin 2	CS	D5	Insight and Action: Diverse Perspectives on Critical Fairness Issues in Testing ~ NCME Committee on Diversity in Testing (CODIT) Featured Session
2:15 PM – 3:45 PM	Belasco	PS	D6	Automatic Item Generation
2:15 PM – 3:45 PM	Plymouth	PS	D7	Developing CDM
2:15 PM – 3:45 PM	Manhattan	PS	D8	Technical Considerations in Assessing DIF

CS=Coordinated Session • EB= Electronic Board Session
IS= Invited Session • PS= Paper Session • TS=Training Session

Time	Room	Type	ID	Title
2:15 PM – 3:45 PM	Ambassador II	CS	D9	Exploring Growth: Methods and Applications
2:15 PM – 3:45 PM	Gershwin I	EB	D10	Electronic Board Session 2
4:05 PM – 6:05 PM	Ambassador III	IS	E1	Measurement Problems 1– A look back to help us look ahead
4:05 PM – 6:05 PM	Broadway I	CS	E2	Measuring Essay Writing Competency in Europe using Human and Automated Scoring
4:05 PM – 6:05 PM	Broadway II	CS	E3	Considerations for Best Practices in Scale Development
4:05 PM – 6:05 PM	Broadway III	CS	E4	Towards Understanding the Facilitators and Inhibitors in Writing Tasks Containing Multimedia-Enhanced Stimuli
4:05 PM – 6:05 PM	Gershwin 2	PS	E5	Detecting Bad Things: Research on Cheating
4:05 PM – 6:05 PM	Belasco	PS	E6	Application and Evaluation of DCM
4:05 PM – 6:05 PM	Plymouth	PS	E7	Investigating Fit
4:05 PM – 6:05 PM	Manhattan	PS	E8	New Research on Multidimensional IRT
4:05 PM – 6:05 PM	Ambassador II	CS	E9	Fairness in Testing ELs and ELs with Disabilities: Research, Implementation, and Policy
6:30 PM – 8:00 PM	Majestic I & II			NCME and Division D Reception
Sunday, April 15				
8:00 AM – 10:00 AM	Broadway I/II/III			NCME Breakfast and Business Meeting
10:35 AM – 12:05 PM	Majestic II	IS	F1	The Positive Impact of Assessment
10:35 AM – 12:05 PM	Ambassador III	PS	F2	Technology-Based Asssesment: Tests, Items, and Methods
10:35 AM – 12:05 PM	Melville	PS	F3	New Directions for Multilevel Models
10:35 AM – 12:05 PM	Majestic I	CS	F4	Students' Use of Response Time, Testing Behavior, and Performance in Digitally-Based Assessments
10:35 AM – 12:05 PM	Gershwin II	CS	F5	Using an Assessment Use Argument in developing, using, and justifying K-12 assessments
10:35 AM – 12:05 PM	Belasco	PS	F6	Issues in Linking and Equating
10:35 AM – 12:05 PM	Plymouth	CS	F7	Exploring Properties, Issues, and Solutions with Estimating Student- and Aggregate-Level Growth Measures
10:35 AM – 12:05 PM	Manhattan	CS	F8	Item Difficulty Modeling: Lessons Learned and Future Directions
10:35 AM – 12:05 PM	Ambassador II	CS	F9	Boundary-pushing innovations in the assessment of English language learners, co-sponsored with AERA-IAEA
2:45 PM – 4:15 PM	Majestic II	IS	G1	Measurement Problems 2– A look back to help us look ahead

CS=Coordinated Session • EB= Electronic Board Session
IS= Invited Session • PS= Paper Session • TS=Training Session

Time	Room	Type	ID	Title
2:45 PM – 4:15 PM	Ambassador III	CS	G2	Tackling practical issues in small sample scaling and equating
2:45 PM – 4:15 PM	Melville	CS	G3	Using Repeater Data to Inspect Quality and Security in Continuous Mode Testing
2:45 PM – 4:15 PM	Majestic I	CS	G4	Assessments of Collaborative Problem Solving and Implications for PISA 2015
2:45 PM – 4:15 PM	Gershwin II	PS	G5	Reimagining Adaptive Testing
2:45 PM – 4:15 PM	Belasco	PS	G6	Approaches to Decisions/Classification
2:45 PM – 4:15 PM	Plymouth	PS	G7	Where Learning and Measurement Meet
2:45 PM – 4:15 PM	Manhattan	PS	G8	Statistical Approaches to Improving Validity
2:45 PM – 4:15 PM	Ambassador II	CS	G9	Assessing mathematical thinking using learning progressions
4:35 PM – 6:05 PM	Majestic II	IS	H1	Award-Winning Research from the 2018 NCME Award Recipients
4:35 PM – 6:05 PM	Ambassador III	CS	H2	The Big Five (Sources of Validity Evidence): Illustrations of Validation Practices
4:35 PM – 6:05 PM	Melville	CS	H3	Dimensionality as it Relates to Primary Latent Factors, Sub-scores, and Item Parcels
4:35 PM – 6:05 PM	Majestic I	IS	H4	New Developments in the Assessment Practice at the National Center for Assessment
4:35 PM – 6:05 PM	Gershwin II	PS	H5	Diving into Data with Response Process Research
4:35 PM – 6:05 PM	Belasco	PS	H6	Modeling Response Times
4:35 PM – 6:05 PM	Plymouth	PS	H7	Scoring Simulations, Performance Tasks, and Polytomous Items
4:35 PM – 6:05 PM	Manhattan	PS	H8	IRT with non-traditional constructs
4:35 PM – 6:05 PM	Ambassador II	CS	H9	Communicating Complex Psychometric Information to Teachers, Parents, and Other Less Technical Audiences
4:35 PM – 6:05 PM	Gershwin I	EB	H10	Electronic Board Session 3
6:30 p.m.–8:00 p.m.	Broadway I/II			President's Reception
Monday, April 16				
5:45 AM – 7:00 AM				NCME Fitness Run/Walk
8:15 AM – 10:15 AM	Broadway I	CS	I2	Mapping state proficiency standards to the NAEP scale: New methods, new results
8:15 AM – 10:15 AM	Broadway II	CS	I3	Emergent Themes from the Development of NGSS-Aligned Summative Science Assessments
8:15 AM – 10:15 AM	Broadway III	CS	I4	Diagnosis and Feedback in Learning and Assessment Systems
8:15 AM – 10:15 AM	Ambassador III	PS	I5	Bayesian Applications
8:15 AM – 10:15 AM	Majestic I	PS	I6	Automated Scoring

CS=Coordinated Session • EB= Electronic Board Session
IS= Invited Session • PS= Paper Session • TS=Training Session

Time	Room	Type	ID	Title
8:15 AM – 10:15 AM	Plymouth	PS	I7	Advances in Communicating Results
8:15 AM – 10:15 AM	Manhattan	CS	I8	Measurement Challenges in On-going Testing Environment: Potential Solutions
8:15 AM – 10:15 AM	Ambassador II	PS	I9	Approaches to Assembly and Administration of Adaptive Tests
10:35 AM – 12:05 PM	Majestic II	IS	J1	We Can Do This: Communicating Information from Educational Assessments
10:35 AM – 12:05 PM	Broadway I	CS	J2	New Insights on Survey Questionnaire Context Effects from Multiple Large-Scale Assessments
10:35 AM – 12:05 PM	Broadway II	CS	J3	Measuring instruction using classroom artifacts and portfolios: Evidence from four recent studies
10:35 AM – 12:05 PM	Broadway III	CS	J4	Innovative Approaches to Standard Setting: Responding to a Changing Assessment Environment
10:35 AM – 12:05 PM	Ambassador III	PS	J5	Emerging Research on the Adaptation of Adaptive Tests
10:35 AM – 12:05 PM	Majestic I	PS	J6	Advances in Estimation of DCM
10:35 AM – 12:05 PM	Plymouth	PS	J7	Exploring Speededness: Detection and Impact
10:35 AM – 12:05 PM	Manhattan	PS	J8	Scoring with Multiple Categories
10:35 AM – 12:05 PM	Ambassador II	PS	J9	Test Score Use, Stakeholder Perceptions, and Evidence of Consequences
10:35 AM – 12:05 PM	Gershwin I	EB	J10	GSIC Graduate Student Poster Session 2
12:00 PM – 2:00 PM	Minskoff			Past Presidents Luncheon
12:25 PM – 1:55 PM	Majestic II	IS	K1	Measurement Problems 3 – A look back to help us look ahead
12:25 PM – 1:55 PM	Broadway I	CS	K2	Validity and Diversity Challenges in Post-Secondary Admissions
12:25 PM – 1:55 PM	Broadway II	CS	K3	Score Reporting for High-Stakes Certification and Licensing Programs
12:25 PM – 1:55 PM	Broadway III	CS	K4	Peeking into Student Writing Behaviors in NAEP: Why and How
12:25 PM – 1:55 PM	Ambassador III	PS	K5	Issues in Growth Modeling
12:25 PM – 1:55 PM	Majestic I	PS	K6	Proficiency Estimation
12:25 PM – 1:55 PM	Plymouth	PS	K7	Processes and Considerations in Adaptive Test Assembly
12:25 PM – 1:55 PM	Manhattan	PS	K8	New Directions in Detecting DIF
12:25 PM – 1:55 PM	Ambassador II	CS	K9	Examining Standard Errors for NAEP Group-score Comparisons Across Years and Digital Transition
12:25 PM – 1:55 PM	Gershwin I	EB	K10	Electronic Board Session 4

Time	Room	Type	ID	Title
2:15 PM – 3:45 PM	Majestic II	CS	L1	Testing in the professions: Credentialing policies and practice
2:15 PM – 3:45 PM	Broadway I	CS	L2	Improving Human Rating
2:15 PM – 3:45 PM	Broadway II	CS	L3	Within and between-high school measurement challenges in college admission
2:15 PM – 3:45 PM	Broadway III	CS	L4	Exploring the Potential Impact of SEL Assessment on School Practices
2:15 PM – 3:45 PM	Ambassador III	PS	L5	IRT for Next Generation Assessments
2:15 PM – 3:45 PM	Majestic I	PS	L6	Impact of People on Linking and Equating
2:15 PM – 3:45 PM	Plymouth	PS	L7	Learning Progressions: Development and Evaluation
2:15 PM – 3:45 PM	Manhattan	PS	L8	Reliability of Scores and Subscores
2:15 PM – 3:45 PM	Ambassador II	CS	L9	Evaluating Paper and Computer Adaptive Test Score Comparability from Multiple Perspectives
2:15 PM – 3:45 PM	Gershwin I	EB	L10	Electronic Board Session 5
4:00 PM – 7:00 PM	Nederlander			Board of Directors
4:05 PM – 6:05 PM	Majestic II	CS	M1	Procedures for Detecting Aberrant Ex AM–Taking Behavior in the Operational Setting
4:05 PM – 6:05 PM	Broadway I	CS	M2	Advances in IRT Equating: Old Methods and New Tricks
4:05 PM – 6:05 PM	Broadway II	CS	M3	Promises and Challenges with Computerized-Adaptive Testing in K-12 Assessments
4:05 PM – 6:05 PM	Broadway III	CS	M4	Collaborating to Measure Collaboration Skills: Principles, Methodologies, and Lessons Learned
4:05 PM – 6:05 PM	Ambassador III	PS	M5	Fairness in Testing Policies and Practices
4:05 PM – 6:05 PM	Majestic I	CS	M6	Challenges, Issues and Opportunities in Using Response Process Data in Improving Measurement
4:05 PM – 6:05 PM	Plymouth	PS	M7	Applications: Understanding Examinee Performance
4:05 PM – 6:05 PM	Manhattan	PS	M8	Modeling, Mediating, and Explaining DIF
4:05 PM – 6:05 PM	Ambassador II	CS	M9	Maintaining quality assessments in the face of change

GIVE TO THE *NCME MISSION FUND*

As measurement professionals, we recognize the rapidly changing needs of teaching, learning, and assessment in the U.S. and globally. Meeting these changing needs through research-based assessment practices continues to remain a priority for NCME. We also recognize that new measurement professionals and graduate students need opportunities to share their work and learn from others in the measurement community in order to ensure that our profession remains strong in the future.

How can YOU make a difference? Contribute to the NCME Mission Fund!

Your donation will help support

- **Support for graduate students, early career faculty, and early career practitioners in the measurement field.** Funding for activities that expand educational and professional opportunities for newer NCME members, both students and active members.

- **Responding to public perceptions of measurement and testing**. Funding for special initiatives outside of existing NCME activities designed to promote a broader understanding of high quality assessment practices and appropriate test use among diverse groups of assessment stakeholders.

- **Co-Sponsorship among NCME committees or with external agencies or organizations.** Funding that involves members of two or more NCME committees or an NCME committee and an external agency or organization in support of activities larger than any single committee's charge.

Information on how to apply for Mission Fund support will be released soon.

National Council on Measurement in Education is very grateful to the following organizations for their generous financial support of our 2018 Annual Meeting

National Council on Measurement in Education
100 North 20th Street, Suite 400, Philadelphia, PA 19103 (215) 461-6263
http://www.ncme.org/