NCME national council on measurement in education

National Council on Measurement in Education

Advancing Large Scale and Classroom Assessment through Research and Practice

2017 Training Sessions

April 26-27

2017 Annual Meeting

April 28-30

San Antonio Marriott Rivercenter San Antonio, TX

#NCME17

Welcome from the Program Chairs

Welcome to the 2017 NCME conference! We are very pleased to have put together this program with the theme, *Advancing Large Scale and Classroom Assessment through Research and Practice*. Below we've presented a few highlights from this year's program.

We have many great sessions related to our theme, including a plenary invited session entitled *Classroom Assessment: Promises, Perils, and Next Steps for Moving Forward* (10:35 AM on Friday, April 28th) chaired by **Jim McMillan**, and featuring NCME president **Mark Wilson**, among other scholars from the field. At 10:35 on Sunday, April 30th, **Bonnie Strykowski**, Vice President of the National Association of Assessment Directors, will lead the session, *Measuring Creativity from Classrooms to Large Scale Assessments: Views from Practice to Research and Development of Assessments*.

During the very first and last NCME time slots, we'll be featuring moderated panel discussion sessions. NCME Board Member **Kristen Huff** will lead a panel discussion of *The Ocean of Data from Classroom EdTech: Are Psychometricians Ready?* at 8:15 on Friday, April 28th. At 4:05 on Sunday, April 30th, we have *The Impact of Accessibility Technology on the Validity of Score Interpretations from Large-Scale and Classroom Assessments*, led by **Anne Davidson**, the chair of NCME's Diversity Committee.

We will be honoring many distinguished members of the measurement field this year. **Linda Cook** is being honored with the NCME Career Award; her presentation, *Testing Individuals with Disabilities: What Constitutes Fairness?* will be held on Saturday, April 29th at 10:35 AM. We will also be having a special session for other recent NCME awardees (12:25 on Friday, April 28th) to present their award-winning research. That session will be followed by a reception to honor the recipients. Lastly, a special invited session honoring the career and contributions of **Ben Wright** will be held at 4:05 on Saturday, April 29th.

We have a fully-loaded program which we hope you'll agree represents both the breadth and depth of the measurement discipline! We've worked hard to ensure that every time slot represents a wide array of topics, including ever-present issues such as standard setting, differential item functioning, and reliability, but also up-and-coming topics, such as test security, automated scoring, and diagnostic classification modeling. As in some previous years, our individual paper sessions don't include discussants; we decided to go in this direction to maximize our acceptance rate for high-quality submissions. This year we were able to achieve an overall 78% acceptance rate among individual and coordinated paper session submissions, but still maintain the very high quality the NCME membership has come to expect.

We are confident that you'll enjoy the 2017 NCME annual meeting program; it would be nothing without participation from esteemed members such as you!

Lydia Liu and Billy Skorupski
2017 NCME Annual Meeting Co-Chairs

Table of Contents

Training Sessions

Program

Advancing Large Scale and
Classroom Assessment through
Research and Practice

NCME Officers

President	Mark Wilson *UC Berkeley, Berkeley, CA*
President Elect	Randy Bennett *ETS, Princeton, NJ*
Past President	Richard J. Patz *ACT, Monterey, CA*

NCME Directors

Luz Bay
The College Board, Dover, NH

Derek Briggs
University of Colorado, Boulder, CO

Kristen Huff
Curriculum Associates, Brooklyn NY

Won-Chan Lee
University of Iowa, Iowa City, IA

Walter Way
The College Board, Pinehurst, NC

C Dale Whittington
Shaker Heights (OH) Public Schools, Shaker Heights, OH

Editors

Journal of Educational Measurement	Jimmy de la Torre *Rutgers, The State University of NJ, New Brunswick, NJ*
Educational Measurement Issues and Practice	Dr. Howard Everson *SRI International, Menlo Park, CA*
NCME Newsletter	Heather M. Buzick *ETS, Princeton, NJ*
Website Content Editor	Brett Foley *Alpine Testing Solutions, Denton, NE*

2017 Annual Meeting Chairs

Annual Meeting Program Chairs	William Skorupski *University of Kansas, Lawrence, KS* Lydia Liu *ETS, Princeton, NJ*
Graduate Student Issues Committee Chair	Brian Leventhal *University of Pittsburgh, Pittsburgh, PA*
Training and Development Committee Chair	Sun-Joo Cho *Peabody College of Vanderbilt University, Nashville, TN*
Fitness Run/Walk Directors	Katherine Furgol Castellano *ETS, San Francisco, CA* Jill R. van den Heuvel *Alpine Testing Solutions, Hatfield, PA*

NCME Information Desk

The NCME Information desk is located on the Meeting Room Level in the Marriott Rivercenter. Stop by to pick up your badge, program book and ribbon, as well as your bib number and t-shirt for the fun run and walk. It will be open at the following times::

Wednesday, April 26	7:30 AM – 4:30 PM
Thursday, April 27	7:30 AM - 4:30 PM
Friday, April 28	8:00 AM – 4:30 PM
Saturday, April 29	10:00 AM – 4:30 PM
Sunday, April 30	8:00 AM – 1:00 PM

Proposal Reviewers

David Abrams
Beyza Aksu-Dunya
Usama Ali
Allison Ames
Randy Bennett
Yanhong Bian
Dan Bolt
Bob Brennan
Brent Bridgeman
Jing Chen
Yi-Ling Chiang
Greg Cizek
Stephen Cubbellotti
Shenghai Dai
William Dardick
Anne Davidson
Qi Diao
John Donoghue
Holmes Finch
Joe Fitzpatrick
EdithAurora Graf
Lin Gu
Lixiong Gu
Kell Harrison
Qiwei He
Bo Hu
Corinne Huggins-Manley
Kyoko Ito
Eunice Eunhee Jang
Yanlin Jiang
Evelyn Johnson
Natalie Jorion
Priya Kannan
Lisa Keller
David Klieger
Tim Konold
Meghan Kuhfeld
Quinn Lathrop
HyeSun Lee
Roy Levy
Chen Li
Guangming Ling
Peiyan Liu
Samuel Livingston
Tanya Longabach
Jaime Malatesta
Ki Matlock
Dan McCaffrey
Hyeonjoo Oh
Maria Elena Oliveri
Andreas Oranje
Jesse Pace
Frank Padellaro
Tianshu Pan
Lu Qin
Heather Rickels
Sam Rikoon
Joseph Rios
Salvador Rivas
Katrina Roohr
Andre Rupp
Matthew Schultz
Lietta Scott
Charles Secolsky
Carl Setzer
Can Shao
Sandip Sinharay
Rabia Esma Sipahi Akbas
William Skorupski
Elizabeth Stone
Hariharan Swaminathan
Jonathan Templin
Anne Traynor
Shiyu Wang
Ting Wang
Jonathan Weeks
Craig Wells
Cathy Wendler
Amanda Wolkowitz
Adam Wyse
Tao Xin
Levent Yakar
Duanli Yan
Hanwook Yoo
Mengxiao Zhu
Cengiz Zopluoglu

Graduate Student Abstract Reviewers

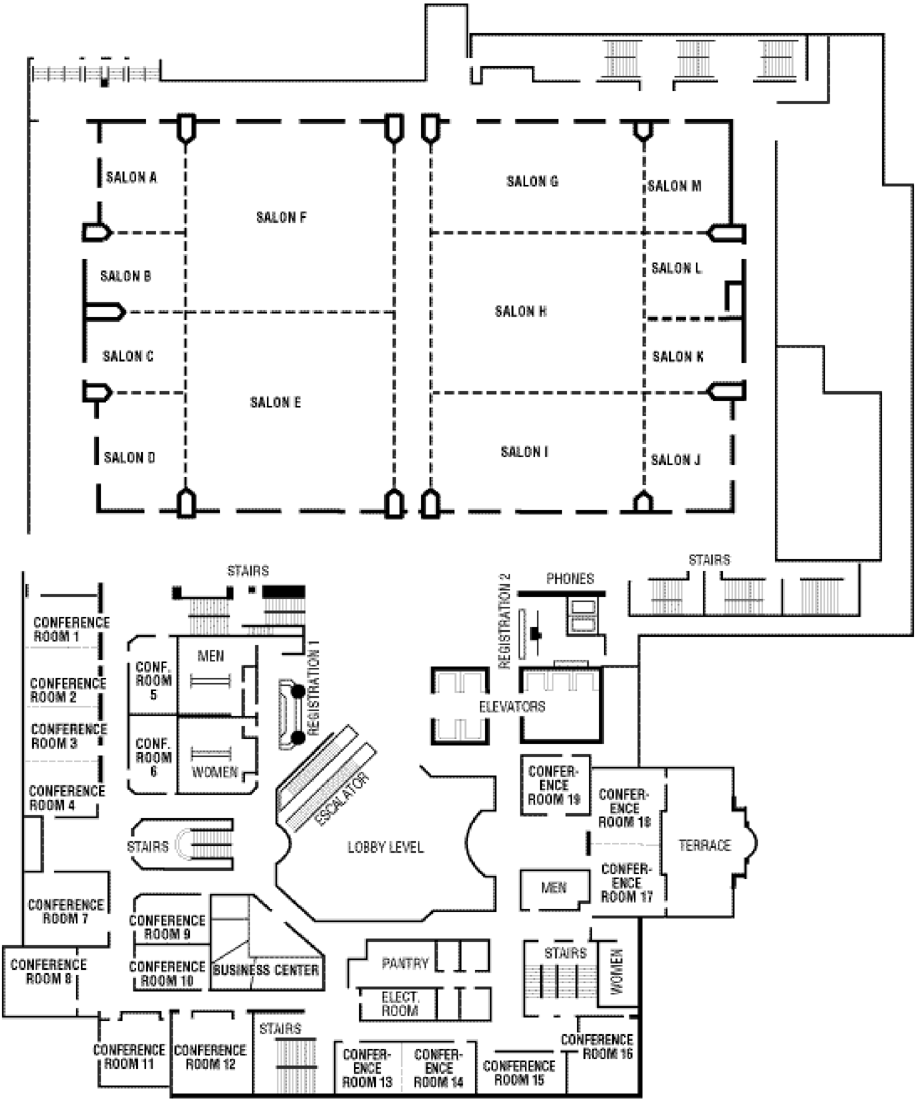
Abeer Alamri
Ezgi Ayturk
Yu Bao
Masha Bertling
Genine Blue
Liuhan Cai
Rajendra Chattergoon
Brandom Craig
Dries Debeer
Sien Deng
Yuyu Fan
Brittany Flanery
Kelly Foelber
Andrew Iverson
Unhee Ju
Kyung Yong Kim
Isaac Li
Chimuma Lilian
Lida Lin
Ye Lin
Ren Liu
Wenchao Ma
James Mason
Michael Mitchell
Sarah Newton
Diep Nguyen
Mary Norris
Yuxi Qiu
Logan Rome
Kevin Carl Santos
Jordan Sparks
Kim Trang
Stephanie Underhill
Jue Wang
Qing Xie
SUJIN Yang
Xin Yuan
Jiahui Zhang
Ya Zhang
Xiaying Zheng

Future Annual Meeting

2018 Annual Meeting
April 12-16
New York, NY, USA

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

Marriott Rivercenter Meeting Room Floor Plans



Pre-Conference Training Sessions

The 2017 Pre-Conference Training Sessions will be held at the Marriott Rivercenter on Wednesday, April 26 and Thursday, April 27. 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 afternoon 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 Marriott Rivercenter for those workshops that still have availability.

Please note that internet connectivity will not be available for most training sessions and, where applicable, participants should download the software required prior to the training sessions. Internet connectivity will be available for a few selected sessions that have pre paid an additional fee.

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

For those graduate students who attend training sessions, we are offering a rebate. In order to receive the rebate, **you must fill out the evaluation form completely** and turn it in. Your full name, address and email address must also be included in order to receive the rebate.

Wednesday, April 26, 2017
8:00 AM–12:00 PM, Salon K, Meeting Room Level, Training Session, AA

Vertical Scaling Methodologies, Applications, and Research

Ye Tong, Pearson
Michael Kolen, University of Iowa

Vertical scaling refers to the process of placing scores on tests that measure similar domains but at different educational levels onto a common scale. Development of vertical scales can help facilitate interpretations of students' achievement from year to year, especially when there is good content alignment between tests of different levels. With many states adopting the common core state standards, there has been a renewed interest in developing vertical scales. The common core state standards are much better vertically aligned across grades and offer a unique content foundation for the development of a vertical scale and a great stage for rethinking on the growth measures towards college readiness. In this training session, the instructors will provide detailed steps for various vertical scaling methodologies, along with examples using both real and synthetic data. The instructors will also provide some examples and discuss the benefits and challenges encountered by various test developers when building vertical scales. Hands-on exercises and interpretations of established vertical scales will also be included.

Wednesday, April 26, 2017
8:00 AM–12:00 PM, Salon L, Meeting Room Level, Training Session, BB

An introduction to Linking and Equating in R

Anthony Albano, University of Nebraska-Lincoln
Jonathan Weeks, Educational Testing Service

The statistical environment R provides open-source access to a variety of psychometric tools, including linking and equating methods for converting scores from multiple test forms to a common measurement scale. This training session introduces participants to observed-score and item response theory linking and equating through interactive lectures and exercises involving analysis of real data in R. Participants will learn to prepare data, conduct linking and equating, and visualize, summarize, and evaluate results. Students, researchers, and practitioners are invited to participate, and should bring their own computers with R installed. A background in introductory statistics and experience using R are recommended but not required.

Wednesday, April 26, 2017
8:00 AM–12:00 PM, Salon M, Meeting Room Level, Training Session, CC

Rubrics for Classroom Assessment: Perils of Practice and How to Avoid Them

Heidi Andrade, University at Albany--SUNY

Rubrics are ubiquitous, and have the potential to both measure and promote student learning. Like any measurement tool, however, the design and implementation of a rubric requires technical knowledge and skill that many educators do not have. Consequently, many rubrics used by teachers are of low quality, not appropriate for use outside a standardized testing context, and/or incorrectly scored. In addition, rubrics are too infrequently used to scaffold the kinds of formative assessment that have been shown to promote learning – but only when the rubrics are of high quality. This half-day, interactive workshop will use research and examples from real classrooms to highlight the promise and pitfalls of rubrics for summative and formative classroom assessment. The learning objectives for the session are to understand:

- The characteristics of a rubric that is appropriate for classroom use
- Common flaws in rubric design and use, and how to avoid them
- The ways in which high-quality rubrics can promote learning via formative assessment
- The features of effective, rubric-referenced peer and self-assessment

The session is appropriate for anyone interested in understanding classroom assessment in order to develop a balanced, comprehensive model of educational assessment. Laptops are not required.

Wednesday, April 26, 2017
8:00 AM–5:00 PM, Salon A, Meeting Room Level, Training Session, DD

Bayesian Networks in Educational Assessment (Book by Springer)

Duanli Yan, Educational Testing Service
Russell Almond, Florida State University
Roy Levy, Arizona State University
Diego Zapata-Rivera, ETS
Robert Mislevy, ETS

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

Wednesday, April 26, 2017
8:00 AM–5:00 PM, Salon B, Meeting Room Level, Training Session, EE

Shadow-Test Approach to Adaptive Testing

Wim van der Linden, Pacific Metrics Corporation
Michelle Barrett, Pacific Metrics Corporation

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.

Wednesday, April 26, 2017
8:00 AM–5:00 PM, Salon C, Meeting Room Level, Training Session, FF

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

Jimmy de la Torre, The University of Hong Kong
Wenchao Ma, Rutgers, The State University of New Jersey

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 psychometric theories. 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, and to download and install the package in advance.

Wednesday, April 26, 2017
8:00 AM–5:00 PM, Salon D, Meeting Room Level, Training Session, GG

Conceptual Frameworks for Aligning Items to ALDs to Enhance Validity Arguments

Christina Schneider, NWEA
Steve Ferrara, Measured Progress

Standards-based achievement tests pursue two goals: (a) Classify students into achievement levels that enable valid inferences about student knowledge and skill; and (b) measure growth toward proficiency. Explicating how the complexity of knowledge and skills in achievement level descriptors (ALDs) are intended to differ, and how that complexity is related to empirical item difficulty—during assessment design, rather than after development—is critical to those goals. And it is consistent with principled approaches to test design, development, and implementation. Experts in test development and standard setting will experiment with emerging practices in developing achievement level descriptors (ALDs) and writing items aligned to those ALDs to support test score inferences and standard setting judgements. Participants will analyze ALDs using a new conceptual framework, item response demand frameworks that support writing items aligned to ALDs, and items from large scale assessments and their alignment with ALDs. Participants will also apply a framework for writing ALDs (Egan, Schneider, & Ferrara (2012), item demands codes (Ferrara, Svetina, Skucha, & Davidson [2011]; Schneider, Huff, Egan, Gaines & Ferrara [2013]), as precursor supports for the ID Matching standard setting method (Ferrara & Lewis, 2012) to support validity arguments for standards-based assessments. Participants should bring their laptops.

Wednesday, April 26, 2017
1:00 PM–5:00 PM, Salon K, Meeting Room Level, Training Session, HH

The History of Educational Measurement in America: Origins to 1950

Michael Bunch, Measurement Incorporated
Michael Beck, BETA
Brian Clauser, National Board of Medical Examiners
Michelle Croft, ACT

Educational measurement professionals, while thoroughly versed in modern theory and practice, may lack a clear understanding of how theory evolved over time and how practice was shaped by forces outside our field. This half-day training session provides an overview of the foundations of educational measurement in the United States, from early influences to the publication of the first edition of Educational Measurement by E. F. Lindquist. Its purpose is to provide historical perspective to the current theories, practices, and policies associated with educational measurement. It consists of four modules and a Q&A session. Module 1 addresses the call for standardization of student assessment, starting in the 1840s. Module 2 focuses on the social/political/legal milieu in which this standardization emerged in the latter half of the 19th and first half of the 20th centuries. Module 3 traces the development of measurement theory to 1950. Module 4 deals with the rise of a professional class of educational measurement specialists and their influence on the field. Participants will have an opportunity to interact with the presenters at the end of each module and during a final question and answer session.

Wednesday, April 26, 2017
1:00 PM–5:00 PM, Salon L, Meeting Room Level, Training Session, II

Landing Your Dream Job for Graduate Students

Deborah Harris, ACT, Inc.
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. It will concentrate on what to do now while they are still in school to best prepare for a job (including finding a dissertation topic, selecting a committee, maximizing experiences while still a student with networking, internships, and volunteering, and providing suggestions to the questions regarding what types of coursework an employer looks for, and what would make a good job talk), how to locate, interview for, and obtain a job (including how to find where jobs are, how to apply for jobs --targeting cover letters, references, and resumes), what to expect in the interview process (including job talks, questions to ask, and negotiating an offer), and what's next after they have started their first post PhD job (including adjusting to the environment, establishing a career path, publishing, finding mentors, balancing work and life, and becoming active in the profession). The session is interactive, and geared to addressing the participants' questions during the session. Resource materials are provided on all relevant topics.

Wednesday, April 26, 2017
1:00 PM–5:00 PM, Salon M, Meeting Room Level, Training Session, JJ

Data Rich, Information Poor: Navigating Data Use in a Balanced Assessment System

Caroline Wylie, Educational Testing Service
Christine Lyon, Educational Testing Service

With continued growth in balanced assessment systems, there are volumes of data available to all stakeholders in K-12 school systems. Given concerns about the amount of student testing, there is a desire to maximize the value of each data point. Consideration of assessment cycles and grain-size (Wiliam, 2006; Shavelson et al., 2008) can support appropriate interpretation and use of different types of data. Different types of data can be used at across the year for informed decision making (e.g., using summative assessment formatively for curriculum review and unit planning, using interim assessment data to confirm or adjust long range planning, using formative assessment for short-cycle adjustments). This half-day workshop will engage participants in discussion about appropriate decision-making across components of a balanced assessment system. The learning goals for the session are for participants to understand: Appropriate uses of summative, interim and formative assessment information How to support the development of strategies for the appropriate use of assessment data by schools and districts How to create more robust PLC discussions around the use of assessment data The session is intended for a wide range of participants including state, district and or school-level staff with responsibilities for assessment. Laptops are not required.

Wednesday, April 26, 2017
1:00 PM–5:00 PM, Conference Room 12, Meeting Room Level, Training Session, KK

An introduction to R for quantitative methods

Brian Habing, University of South Carolina
Jessalyn Smith, DRC

The free statistical package R has become a favorite of researchers over the past decade – and is now increasingly used in operational work and teaching methods courses at all levels. With you working along through each step with us, this course will cover some of the most useful aspects of R, with a focus on using it for quantitate methods. The session begins with an introduction to the language (including data management, graphing, and statistical analysis). The second portion covers R's implementation of the quantitative methods and a sample of commonly used psychometric methods (including an overview of appropriate packages and some useful custom made functions). The final portion examines R's use for conducting customized functions and simulations. This will also include an introduction to popular psychometric packages. This course is designed for current or future practitioners who use quantitative methods in his/her own work, research, or teaching methods but have little to no previous experience with R. Participants must bring their own laptop computer; all required software will be provided.

Wednesday, April 26, 2017
1:00 PM–5:00 PM, Conference Room 1&2, Meeting Room Level, Training Session, LL

Analyzing NAEP Data Using Plausible Values and Marginal Estimation with AM

Emmanuel Sikali, National Center for Education Statistics
Young Yee Kim, American Institutes 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 model. 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.

Thursday, April 27, 2017
8:00 AM–12:00 PM, Salon L, Meeting Room Level, Training Session, MM

Evidenced-Centered Design and Computational Psychometrics Solution for Game/Simulation-based Assessments

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

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).

Thursday, April 27, 2017
8:00 AM–12:00 PM, Salon M, Meeting Room Level, Training Session, NN

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

Katie Brien, Pearson
Ye Tong, Pearson

Online assessments provide a great opportunity for students to interact with content in a more dynamic way, but moving away from paper or to both paper and online assessments also 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. The most common online item types and results will be used to increase application across states and platforms. Participants will receive instruction around common practices and challenges faced when developing technology-enhanced items (TEIs). Hands-on experience will also be incorporated in the training session around making decisions about assessment formats, item types, and scoring associated with these item types.

Thursday, April 27, 2017
8:00 AM–5:00 PM, Salon A, Meeting Room Level, Training Session, OO

Diagnostic Classification Models: Theory, Methods, and Applications

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. This workshop provides a semi-technical, hands-on introduction to the terms, techniques, and methods used for diagnosing what students know for purposes of guiding decisions regarding students' instructional needs. Upon completion of the workshop, participants will be able to understand the rationale and motivation for using diagnostic measurement methods. Participants will also be able to understand how to design a diagnostic assessment and to interpret information obtained from the diagnostic models. The target audience members are educational researchers and practitioners. This session is appropriate for graduate students, researchers, and practitioners at the emerging or experienced level. Participants need only a basic knowledge of statistics and psychometrics to enroll. The instructors will alternate teaching/facilitating sections. Much of the content will be delivered through lecture, and content will be reinforced using hands-on activities embedded throughout the sections. Participants are encouraged but not required to bring a laptop. Material will be presented at a technical level when necessary for understanding the models and applying them responsibly. Instructors will encourage audience participation through questions and allow time for interactive discussions.

Thursday, April 27, 2017
8:00 AM–5:00 PM, Salon B, Meeting Room Level, Training Session, PP

Interpersonal and Intrapersonal Skills Assessment: Design, Development, Scoring, and Reporting

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

This workshop will provide training, discussion, and hands-on experience in developing methods for assessing, scoring, and reporting on students' social-emotional and self-management or character skills. Workshop will focus on (a) reviewing the kinds of character skills most important to assess based on current research; (b) standard and innovative methods for assessing character skills, including self-, peer-, teacher-, and parent- rating-scale reports, forced-choice (rankings), anchoring vignettes, and situational judgment methods; (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) validation strategies, including the development of rubrics and behaviorally anchored rating scales, and correlations with external variables; (f) the use of anchors in longitudinal growth studies, (g) reliability from classical test theory (alpha, test-retest), item-response theory, and generalizability theory; and (h) reporting issues. These topics will be covered throughout the workshop where appropriate, workshop sessions will tend to be organized around item types (e.g., forced-choice, anchoring vignettes). Examples will be drawn from various assessments, including PISA, NAEP, SuccessNavigator, FACETS, and others. The workshop is designed for a broad audience of assessment developers, analysts, and psychometricians, working in either applied or research settings.

Thursday, April 27, 2017
8:00 AM–5:00 PM, Salon C, Meeting Room Level, Training Session, QQ

Bayesian Estimation of Item Response Theory Model Parameters Using OpenBUGS and Stan

Hong Jiao, University of Maryland, College Park
Yong Luo, National Center for Assessment in Higher Education
Kaiwen Man, University of Maryland, College Park
Dandan Liao, University of Maryland

This session will provide audience with systematic training on Bayesian estimation of standard and extended item response theory (IRT) models using software programs, OpenBUGS and Stan. It covers the basics of these two software programs. The estimation of model parameters for unidimensional dichotomous and polytomous IRT models, multidimensional including testlet models, and multilevel IRT models will be illustrated and demonstrated using both OpenBUGS and Stan. Further the advantages and disadvantages of using each software program will be discussed. This session consists of lecture, demonstration, and hands-on activities of running OpenBUGS and Stan. It is intended for intermediate and advanced graduate students, researchers, and practitioners who are interested in learning the basics and advanced topics related to IRT model parameter estimation using two Bayesian estimation software programs OpenBUGS and Stan. It is expected the audience will have some basic knowledge of the Bayesian theory, but not required. Attendees will bring their own laptop and download the software programs free online. It is expected that attendees will master the basics of writing OpenBUGS and Stan codes in running standard and extended IRT models; further they can develop OpenBUGS and Stan codes for new IRT models for their own research and psychometric modeling.

Thursday, April 27, 2017
8:00 AM–5:00 PM, Salon D, Meeting Room Level, Training Session, RR

An Introduction to Hierarchical Rater Models for the Analysis of Ratings

Jodi Casabianca, Educational Testing Service
Ricardo Nieto, The University of Texas at Austin

Rater effects in education testing and research can impact the quality of scores in constructed response and performance assessments. The hierarchical rater model (HRM) is a multilevel item response theory model for multiple ratings of behavior/performance that yields estimates of latent traits corrected for individual rater bias and variability. The class of HRMs has statistical advantages over other rater models and because of these advantages, recent research has focused on extending this model for multidimensional rubrics and longitudinal assessments. This training session will be presented in four units and aims to instruct those interested in the HRM and these recent extensions. The first unit provides an overview of IRT rater models and the basic parameterization of the HRM. The second and third units introduce the longitudinal and multidimensional HRMs, respectively. The last unit focuses on fitting the HRM and will include an introduction to estimating Bayesian IRT models, and the application of the HRM to empirical and simulated datasets using JAGS and R. Participants will gain hands-on experience fitting the HRM and should bring their laptops with preloaded software. This session is intended for graduate students and professionals that have a solid understanding of IRT models and experience using R.

Thursday, April 27, 2017
8:00 AM–5:00 PM, Salon K, Meeting Room Level, Training Session, SS

A Framework and Platform for the Development of Assessment Literacy

Damian Betebenner, National Center for the Improvement of Educational Assessment
Charles Depascale, National Center for the Improvement of Educational Assessment
Luciana Cancado, University of Wisconsin, Milwaukee
Amy Sharpe, University of Washington
Kelli Ryan, Kent State University

As education measurement takes on greater prominence in the evaluation of students, the districts and schools they attend, and, increasingly, their teachers, there is greater scrutiny of the proper use and misuse of assessment results. Despite the best efforts of measurement professions, even results of the highest technical quality assessments can be subject to misunderstanding and abuse. In recent years, efforts directed at Assessment Literacy have become much more wide spread as our field attempts to assist users of assessments to understand and sensibly apply their assessment results. The training session (reminiscent of a hack-a-thon) will contribute to that effort by assisting users in the development of interactive assessment literacy modules that are (i) browser based, (ii) immediately publishable/available via the web, and (iii) suitable for collaboration/sharing with others.

Thursday, April 27, 2017
1:00 PM–5:00 PM, Salon L, Meeting Room Level, Training Session, TT

Computerized Multistage Adaptive Testing: Theory and Applications

Duanli Yan, Educational Testing Service
Alina von Davier, Educational Testing Service
Kyung Han, GMAC

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.

Thursday, April 27, 2017
1:00 PM–5:00 PM, Salon M, Meeting Room Level, Training Session, UU

Using Visual Displays to Inform Assessment Development and Validation

Brett Foley, Alpine Testing Solutions

The development of an assessment program draws on the expertise of testing professionals for procedural guidance and the knowledge and judgment of subject matter experts (SMEs) who are familiar with the content and testing population of interest. In addition to development, consumers of test results (e.g., students, parents, candidates, policymakers, public), rely on score reports and related documentation to help interpret test scores. In this workshop, we illustrate how visual displays can help inform steps of the test development and validation process, from program design to item writing and review to communicating results through score reporting. Relevant examples of visual displays are provided for various development activities in a range of testing settings (e.g., education, licensure, certification). The presenter will provide step-by-step instruction on how to create the various displays using readily available software. Participants should bring a laptop loaded with Microsoft Excel (2010 version, or later, highly recommended). Panelists will receive flash drives with Excel files and instructions for creating and adapting the visuals discussed in the workshop.

Thursday, April 27, 2017
1:00 PM–5:00 PM, Conference Room 12, Meeting Room Level, Training Session, VV

Evaluating Alignment of Computer Adaptive Assessments

Katerina Schenke, CRESST/University of California, Los Angeles
Deborah La Torre, CRESST/University of California, Los Angeles

With the introduction of college and career ready standards, such as the Common Core, high stakes assessments have gone from mostly fixed, paper-pencil forms, to computer-based adaptive assessments (CAT) that claim to be more reliable and efficient in estimating a student's proficiency. The proposed workshop provides a theoretical and hands-on approach to evaluating CAT delivery systems. The workshop is divided into two sections demonstrating a comprehensive mixed-methods approach. The first section is a step-by-step guide to assessing alignment through the qualitative coding of items. Issues that will be covered include the constructing and training of panels, how to evaluate the alignment of standards, and the rating of complexity with a specific emphasis on depth of knowledge (DOK) alignment. Next, the proposed workshop will cover current thinking and ways to quantitatively evaluate the CAT system by considering blueprint fidelity, item exposure, and reliability and difficulty of test instances. Finally, the presenters will provide an overview of how both qualitative and quantitative methods can be used to address evaluation of CAT delivery systems. The intended audience for this workshop are assessment specialists, school administrators, and researchers working in the area of content alignment. Demonstrations will be shown using R software.

Thursday, April 27, 2017
1:00 PM–5:00 PM, Conference Room 1&2, Meeting Room Level, Training Session, WW

A Visual Introduction to Computerized Adaptive Testing

Yuehmei Chien, Pearson

The training will provide the essential background information on computerized adaptive testing (CAT) with an emphasis on operational CAT algorithm design for different types of assessment including classroom assessment, summative assessment and licensure examination. The multistage CAT, a special case of regular CAT, will also be covered, including stage design and stage panel assembly. The CAT design and theory are presented through an interactive visualization web application for CAT that is developed by the trainer. The training includes lectures, demonstrations, and hands-on activities. 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 web applications designed to help the participants understand different CAT components and concepts and visualize the simulation results. Electronic training materials will be provided via email prior to the conference. Upon completion of the workshop, participants will have deeper understanding of various CAT algorithms, realize the insightful information through examples, configure CAT setting and evaluate the results through simulation, and construct multistage forms using automated assembly tools.

Thursday, April 27, 2017
4:00 PM–7:00 PM, Conference Room 9, Meeting Room Level

NCME Board of Directors Meeting

Members of NCME are invited to attend as observers.

Friday, April 28, 2017
6:30 AM–7:30 AM, Rio Vista Room, Meeting Room Level

Yoga

Please join us for the second NCME Sunrise Yoga. We will start promptly at 6:30 a.m. for one hour at the Marriott Rivercenter. Advance registration required ($10) to reserve your mat. NO EXPERIENCE NECESSARY. Just bring your body and your mind. Namaste.

Friday, April 28, 2017
8:15 AM–10:15 AM, Salon A, Meeting Room Level, Invited Session, A1

The Ocean of Data from Classroom EdTech: Are Psychometricians Ready?

Session Chair: Kristen Huff, Assessment and Research, Curriculum Associates
Session Panelists:
John Behrens, Pearson
Brad Bernatek, Gates Foundation
Derek Briggs, University of Colorado
Kathleen Scalise, University of Oregon
Bi Vuong, Center for Educational Policy Research, Harvard

This invited panel will explore how the influx of data from student use of technology for assessment and learning is shaping - and will continue to shape - measurement theory and practice. The panel will consist of experts from within NCME as well as guests from Harvard's Center for Educational Policy, which evaluates the role of technology in the classroom, and the Gates Foundation, which funds multiple research and development projects to further the advancement of technological innovation in classroom assessment and learning.

Friday, April 28, 2017
8:15 AM–10:15 AM, Conference Room 1&2, Meeting Room Level, Coordinated Session, A2

Challenges in Automated Scoring beyond Features and Models

Session Organizer: Scott Wood, Pacific Metrics Corporation

Automated scoring is the use of computer algorithms to score constructed response items in a way that emulates human scorers. Given the recent publication of large-scale automated scoring studies and freely-available engine code, automated scoring has become more accessible to test developers and users. The purpose of this coordinated session is to sample current automated scoring research that extends beyond the traditional study of features and statistical models. The presenters hope to encourage interaction between automated scoring professionals and other assessment professionals, such as item writers, content experts, writing instructors, test users, test stakeholders, and sales and marketing staff. The Interaction between Automated Scoring and Item Format

The Interaction between Automated Scoring and Item Format
Matthew Schultz, American Institute of Certified Public Accountants; Joshua Stopek, American Institute of Certified Public Accountants; Scott Wood, Pacific Metrics Corporation; Sue Lottridge, Pacific Metrics Corporation; Carlo Morales, Pacific Metrics Corporation

The Automatic Detection of Disturbing Content
Amy Burkhardt, University of Colorado, Boulder; Carlo Morales, Pacific Metrics Corporation

Automatic Scoring and Formative Feedback
Goran Lazendic, Australian Curriculum Assessment and Reporting Authority; Martina Bovell, Australian Curriculum Assessment and Reporting Authority; Sue Lottridge, Pacific Metrics Corporation; Amy Burkhardt, University of Colorado, Boulder

Evaluating Automated Scoring Systems with Validity in Mind: Methodological Design Decisions
Andre Rupp, Educational Testing Service

Media, the Public, and Automated Scoring
Scott Wood, Pacific Metrics Corporation

Friday, April 28, 2017
8:15 AM–10:15 AM, Conference Room 3&4, Meeting Room Level, Coordinated Session, A3

Cognitively Diagnostic Assessment of Middle School Proportional Reasoning: Development and Analysis

Session Organizer: Jimmy de la Torre, The University of Hong Kong
Session Discussant: Jeffrey Douglas, University of Illinois

Despite their increasing popularity, the purported advantages of cognitive diagnosis models (CDMs) have not been fully empirically demonstrated. This is primarily because many CDM applications involve cognitively diagnostic assessments (CDAs) that have been designed to only measure unidimensional constructs. To fill this gap in the literature, this coordinated session will present a middle school proportional reasoning (PR) test designed to be a CDA from its inception. In addition to the test, the session will also illustrate some CDM analyses based on the generalized deterministic input, noisy "and" gate model framework that can be carried out using the PR data. Topics that will be discussed in the session include: steps involved in developing the PR test and the attributes it measures; the advantages of fitting a CDM to an assessment designed to be cognitively diagnostic; evaluation of the quality of the PR test; differential item functioning and person-fit analyses; and converting the PR test to a computerized adaptive test. As a whole, the coordinated session aims to provide empirical evidence of the advantages of CDMs and the rich information that they can offer, and to demonstrate that CDMs are a practically viable tool for carrying out informative and relevant classroom assessment.

Proportional Reasoning Attribute and Test Development
Hartono Tjoe, Penn State Berks; Jimmy de la Torre, The University of Hong Kong

Choosing between CDM and IRT: The Proportional Reasoning Test Case
Wenchao Ma, Rutgers, The State University of New Jersey; Nathan Minchen, Rutgers, The State University of New Jersey

Classification Accuracy and Consistency of the Proportional Reasoning Test
Charles Iaconangelo, Rutgers, The State University of New Jersey; Juan Barrada, Universidad de Zaragoza

Examining DIF in the Proportional Reasoning Test using Various Wald Test Formulations
Likun Hou, Educational Testing Service; Ragip Terzi, Rutgers, The State University of New Jersey

Assessing Person Fit for the Proportional Reasoning Test
Kevin Santos, University of the Philippines- Diliman; Weiling Deng, Educational Testing Service

CD-CAT Implementation of the Proportional Reasoning Assessment
Miguel Sorrel, Universidad Autónoma de Madrid; Hulya Yigit, Rutgers, The State University of New Jersey; Mehmet Kaplan, The Turkish Ministry of National Education

Friday, April 28, 2017
8:15 AM–10:15 AM, Salon B, Meeting Room Level, Coordinated Session, A4

Modeling Learning Progressions and Informing Practice

Session Organizer: Lei Yu, Measured Progress
Session Discussant: Mark Reckase, Michigan State University

Traditional use of vertical scales as a growth measure has been criticized for distorting growth due to construct shift across grade levels and disconnecting inferences made about criterion-referenced growth and norm-referenced metric. A key shift in the common core state standards (CCSS) that started the next generation of assessments is the use of the learning progression-based principles that postulate a staircase of increasing complexity and progressive development of skills as students move from grade to grade. This vertical trajectory of learning defined by a learning progression (LP) captures the essence of learning pathways and supports growth modeling using vertical scales. As new assessments are developed to align with CCSS, LPs become important considerations in test development and scaling. This session provides a thorough and systematic study of modeling LPs using unidimensional and multidimensional IRT, new procedures to evaluate the parameter stability of vertical linking items and the robustness of the Stocking-and-Lord transformation, and creating innovative criterion-referenced student reports to provide information about where students are located on each LP using cognitive maps. The findings from this symposium will provide empirical guidance on modeling LP-based assessments for practitioners and using the information to guide research and practice to improve student learning.

Establishing Learning Progression-based Vertical Scales to Measure Growth
Lei Yu, Measured Progress; Wonsuk Kim, Measured Progress; Jennifer Dunn, Measured Progress

Evaluation of Vertical Linking Items in a Learning Progression context
Unhee Ju, Michigan State University; Jing Jiang, Boston College; Han Yi Kim, Measured Progress

Multidimensional Modeling of Learning Progression-based Vertical Scales
Nina Deng, Measured Progress; Louis Roussos, Measured Progress; Lee LaFond, Measured Progress

Reporting Growth on Learning Progressions
Catherine Taylor, Measured Progress

Friday, April 28, 2017
8:15 AM–10:15 AM, Salon C, Meeting Room Level, Coordinated Session, A5

Large-scale social-emotional skills assessment in K-12

Session Chairs: Patrick Kyllonen, Educational Testing Service; Jinghua Liu, SSATB
Session Discussant: Wayne Camara, ACT

The topic of noncognitive skills has created a buzz in education circles due to growing evidence of their importance in school. This is shown in attention given in popular media, large-scale assessments, and state and national legislation. Measuring noncognitive skills is likely to become an increasingly important topic for the educational measurement community. The purpose of this symposium is to discuss assessment development, use, and findings from five of the most significant projects in the K-12 noncognitive skills domain currently underway. Papers 1 and 4 present findings from two assessments designed to measure social-emotional skills to indicate student readiness and to monitor institution level changes in secondary schools. Paper 2 presents findings from a College Board project to measure noncognitive skills in 8th and 10th grade and their relationship to cognitive skills (PSAT). Paper 3 describes results from California's CORE program, the nation's first attempt at using noncognitive measures for accountability. Paper 5 describes NAEP efforts to add "perseverance, self-control, and need for cognition" to supplement achievement reporting. All papers address issues critical to both researchers and practitioners: What noncognitive skills are crucial, what are the best measurement methods, and how should score uses be validated?

Development of Character Skills Assessment for Secondary Schools
Jinghua Liu, Enrollment Management Association; Kevin Petway, Educational Testing Service; Meghan Brenneman, Educational Testing Service; Cristina Anguiano-Carrasco, Educational Testing Service; Christopher Kurzum, Educational Testing Service

Development of a Noncognitive Assessment for Supporting College Readiness and Success
Carol Barry, College Board; YoungKoung Kim, College Board; Weiwei Cui, College Board; Tim Moses, College Board

Measuring students' social-emotional skills: Evidence from California's CORE districts
Martin West, Harvard Graduate School of Education; Aaron Dow, Harvard Graduate School of Education; Katie Buckley, Transforming Education

Great Rater Debate: Understanding Students' and Teacher's Ratings of Students' SEL Skills
Kevin Petway, Educational Testing Service; Jessica Flake, York University

Noncognitive Survey Questionnaire Indices for NAEP
Jonas Bertling, Educational Testing Service; Ryan Whorton, Educational Testing Service; Jamie Deaton, National Center for Education Statistics

Friday, April 28, 2017
8:15 AM–10:15 AM, Salon D, Meeting Room Level, Paper Session, A6

Test Security and Model Fit

Session Chair: Jeffrey Steedle, Pearson

A Model-Based Market Basket Analysis for Identifying Group Cheating

Jyun-Hong Chen, National Sun Yat-sen University; Hsiu-Yi Chao, National Chung Cheng University

Due to the development of technology, examinees may receive answers from the same source with electronic techniques. Since traditional answer copying detection methods are inapplicable in such circumstance, this study proposes a two-step model-based market basket analysis for detecting group answer copying and proves its efficiency with simulation studies.

Detecting Aberrant Score Patterns on Mixed-Format Tests Using the Modified Caution Index

Jeffrey Steedle, Pearson

The Modified caution index (MCI) can detect aberrant response patterns. A novel polytomous generalization of MCI was applied to simulated data from six assessments. This statistic detected aberrant responding only for lower ability examinees exhibiting high levels of aberrant responding. An IRT-based person fit measure demonstrated greater sensitivity.

Detecting Answer Copying: With or Without Response Times

Heru Widiatmo, ACT, Inc.; Chi-Yu Huang, ACT, Inc.

The copying detection index based on the bivariate lognormal model (van der Linden, 2009) which incorporates response times (RTs) is compared to the ω index model (Wollack, 1997) which does not incorporate RTs. Real data and simulation data are used to discuss the degree of detection consistency from two indices.

Detection of Answer Copying from Multiple Sources

Hsiu-Yi Chao, National Chung Cheng University; Jyun-Hong Chen, National Sun Yat-sen University; Shu-Ying Chen, National Chung Cheng University

Examinees may copy answer from more than one source; however, existing methods were proposed for detecting answer copying for a pair of examinees. This study, thus, proposes a general method for detecting answer copying from multiple sources and proves its efficiency over traditional methods with a simulation study.

Examining the performance of for test security under test speededness

Aijun wang, federation of state boards of physical therapy; Allan Cohen, University of Georgia; Yu Zhang, federation of state boards of physical therapy; Lorin Mueller, federation of state boards of physical therapy

Person-fit indices have been shown to be a useful tool for the detection of item pre-knowledge. They identify unusual response patterns without specifying the reasons. In this study, we examined the performance of the index for identifying examinees with item pre-knowledge when a test is speeded.

Increasing the Robustness of the Deterministic Gated IRT Model to Detect Preknowledge

Carol Eckerly, Alpine Testing Solutions; James Wollack, University of Wisconsin - Madison

This paper introduces a scale purified approach using the Deterministic Gated IRT Model to identify both examinees who have likely benefitted from item preknowledge and compromised items. This method reduces bias in parameter estimates used in the model for classification of examinees as either those with preknowledge or without preknowledge.

Friday, April 28, 2017
8:15 AM–10:15 AM, Salon K, Meeting Room Level, Paper Session, A7

Linking & Equating with IRT

Accuracy of Percent Reaching Standards under IRT True Score Equating

Songbai Lin, Educational Testing Service; Ying Lu, Educational Testing Service

This study examines how much the accuracy of Percent Reaching Standards (PRS), or trend analysis based on PRS statistics, is affected by IRT true score equating. While IRT true score equating equates test forms through true score relationship, justification is needed for comparability of PRS statistics calculated through observed scores

Impact of Anchor Item Drift on Equating Accuracy for Mixed-format Tests

Zhen Li, Government of Newfoundland and Labrador; Haiqin Chen, American Dental Association

This study investigates format effect (FE), item parameter drift (IPD), and their interaction on equating accuracy for mixed-format tests. Separate, concurrent, and fixed parameter calibration methods are compared. Results will provide information on how, when, and to what extent the IPD, FE and their interaction affect equating accuracy.

Impact of Item Parameter Drift on Mixed-Format Tests

Yong He, ACT; Qing Yi, ACT;

Item parameters change over time is defined as item parameter drift (IPD). Previous research examined IPD impact on ability estimates with multiple-choice items. This study extends such studies to a mixed-format test. Calibrated item parameters from a mathematics test are used as the basis of simulation in the study.

Linking in Mixture IRT Models

TTUGBA KARADAVUT, UNIVERSITY OF GEORGIA; ALLAN S COHEN, UNIVERSITY OF GEORGIA

Linking in mixture IRT models typically uses parameter estimates from all latent classes for calculating transformation coefficients. This procedure does not account for estimation errors due to sample size differences between latent classes. Another approach to be studied in this paper is linking each latent class individually.

Scale Transformation Methods to Link a MIRT Calibration to the Item Pool

Tsung-Han Ho, ETS

Two MIRT linking methods that are not associated with the estimation of transformation coefficients and matrices are introduced. The performance of simultaneous linking and parameter-constraint linking is evaluated by the comparison with MIRT Stocking-Lord procedure in terms of the item parameter recovery and the stability of transformation across test conditions.

The Effects of Construct-Shift on Item-level Model-Fit and Unidimensional IRT True-score Equating

Peng Lin, Educational Testing Service; Neil Dorans, Educational Testing Service

This study investigates the effects of construct shift on item-level model fit and IRT true-score equating, for different unidimensional IRT models (1PL vs 2PL) and item calibrations (concurrent vs separate) in single population. Equating results and model fit results will be examined and interpreted.

Friday, April 28, 2017
8:15 AM–10:15 AM, Salon L, Meeting Room Level, Paper Session, A8

Validity Issues and International Assessment

Session Chairs: Jamie Dunlea, British Council; Maria Vasquez-Colina, Florida Atlantic University

An interpretation/use argument for public examinations at 11+ in Trinidad and Tobago

Jerome De Lisle, The University of the West Indies; Janet Ramnanan-Mungroo, The University of the West Indies

An interpretation/use argument (IUA) provides insight into the evidentiary reasoning and justifications for interpretations and uses of test scores from high stakes 11+ public examinations in the Caribbean. Forty-two selected policy claims from 1960 to 2013 were rank ordered and evaluated by stakeholders using Q-methodology.

Fostering Classroom Assessment Knowledge through Dialogue: A Mayan parent perspective

Maria Vasquez-Colina, Florida Atlantic University

This study used focus-groups to interview Mayan parents of K-12 children about their classroom assessment knowledge. Results suggest parents value assessments used for their children, but they were unclear of its purpose. Implications to help parents ask assessment questions and to become more knowledgeable in the educational process are discussed.

Modelling English language learning outcomes from primary school students in rural India

Karen Dunn, British Council; Jamie Dunlea, British Council

This paper describes how analysis of large-scale data has contributed valuable insights to discussions surrounding educational policy for English language education in India. The focus is on the appropriate statistical representation of the complex relationship between first and second language performance and varied student ages at each primary grade level.

Tensions and synergy: validation and quality assurance in a large-scale EFL test

Jamie Dunlea, British Council; Karen Dunn, British Council

This paper explores tensions and synergies in the validation of an international EFL proficiency test within a relatively unexplored space for test development: meeting rigorous technical standards in localized non-certificated test uses. Operational data from a 12-month period were analysed to critically assess assumptions made at pretesting and pre-equating stages

The Applicability of the WJ-III Academic Knowledge Test with Canadian Adult Learners

Charles Hunter, Georgia State University; Hongli Li, Georgia State University; Lee Branum-Martin, Georgia State University; Erin Brown, Georgia State University; Daphne Greenberg, Georgia State University; Jan Frijters, Brock University

A few WJ-III Academic Knowledge items were adapted for Canadian examinees because they are U.S. specific and could be biased against Canadian examinees. DIF analysis reveals some success of the adaption and unexpected DIF for other items. Caution is needed for a test to be used in a different country.

Time-on-task-effects in digital reading: Issues of replicability

Johannes Naumann, Goethe University; Frank Goldhammer, German Institute for International Educational Research (DIPF)

Using PISA 2009 and 2012 Digital Reading Assessment data we show that time-on-task-effects in digital reading might vary due to procedural differences. The pattern of time-on-task-effects being strong in hard items and weak subjects, and low or negative in easy items and strong subjects was only partly replicable across cycles.

Friday, April 28, 2017
8:15 AM–10:15 AM, Salon M, Meeting Room Level, Paper Session, A9

Reliability Issues in IRT

Examining Differential Item Functioning in the Context of Large Scale Assessments
DIMITER DIMITROV, National Center for Assessment

This paper offers an approach to testing for DIF of binary items from both IRT and CFA perspectives in the framework of latent variable modeling. A source code in Mplus is provided to deliver estimates of differential functioning at item and test levels, with confidence intervals obtained via bias-corrected bootstrapping.

Exploring Within-Rater Category Ordering: A Simulation Study using Adjacent-Categories Mokken Scale Analysis
Stefanie Wind, The University of Alabama; Randall Schumacker, The University of Alabama

Molenaar's original polytomous Mokken scaling models are based on cumulative probabilities. Consequently, they cannot be used to empirically verify rating scale category threshold ordering. This study explores the sensitivity of polytomous Mokken models based on adjacent categories to category disordering. Simulation results suggest that these models can detect disordered thresholds.

Response Styles and the Mixed PCM IRT Model: A Simulation
Bruce Austin, Washington State University; Brian French, Washington State University; David Alpizar, Washington State University

This simulation study examined the accuracy of the mixed item response partial credit model for adjusting data for cultural response styles. Acquiescent, mid-point, extreme, and whole-scale responding were simulated with four and five response categories items. Results showed mixed performance but improvement in predictive validity for the 5-category response data.

Scale Transformations with Stabilized Conditional Standard Errors of Measurement for Mixed-Format Tests
Shichao Wang, The University of Iowa; Michael Kolen, The University of Iowa

This study evaluates three raw-to-scale score transformations that can be used to produce stabilized CSEM. The transformations considered are the arcsine transformation along with two recently developed methods, the general variance stabilization and cubic transformation methods. Mixed-format tests containing dichotomous and polytomous items are the focus of the evaluation.

Using POMDPs to Implement an RTI Framework for Early Reading
Umit Tokac, Florida State University

This study compares the POMDP-RTI model with the current-time only-RTI, evaluating the predictive accuracy of each model by checking properly identified model parameters, the quality of the instructional plans produced and the reading levels achieved at the end of the year.

Friday, April 28, 2017
10:35 AM–12:05 PM, Salon EF, Meeting Room Level, Invited Session, B1

Classroom Assessment: Promises, Perils, and Next Steps for Moving Forward

Session Chair: Jim McMillan, Department of Foundations of Education, Virginia Commonwealth University

The purpose of this session is to provide a forum for perspectives, ideas, and discussion about how key research findings in student learning and motivation, in the context of ubiquitous large-scale assessment, can be used by the educational measurement community to promote high quality classroom assessment. While research on classroom assessment has recently accelerated, there remains a need for educational measurement specialists to incorporate what is known from research and theory on learning and motivation, as well as recent developments in large-scale testing that have significant impacts on classroom assessment, to advance teachers' assessment practices that improve as well as document student learning.

Connecting Formative Assessment Practices to Learning Theory
Lorrie Shepard

The Role of Feedback in Assessment
Susan Brookhart

Self-regulated Learning and Classroom Assessment
Heidi Andrade

The Role of Student Effort in Assessment
Steven Wise

On the Relationship Between Large-Scale Assessment and Classroom Assessment
Mark Wilson

Friday, April 28, 2017
12:25 PM–1:55 PM, Salon A, Meeting Room Level, Invited Session, C1

NCME Award Winners

2017 NCME Alicia Cascallar Award for an Outstanding Paper by an Early Career Scholar
Dr. Qiwei He

2017 NCME Annual Award
Dr. Kyung (Chris) Han

2017 NCME Bradley Hanson Award for Contributions to Educational Measurement
Dr. Wenchao Ma and Dr. Jimmy de la Torre

2017 NCME Jason Millman Promising Measurement Scholar Award
Dr. Minjeong Jeon

2017 NCME Brenda H. Loyd Outstanding Dissertation Award
Dr. Megan Kuhfeld

Friday, April 28, 2017
12:25 PM–1:55 PM, Conference Room 1&2, Meeting Room Level, Coordinated Session, C2

Bayesian Developments in Modeling and Prediction for LSA Data with Applications

Session Chair: Jean-Paul Fox, University of Twente
Session Discussant: Matthias von Davier, Educational Testing Service

International large-scale assessments (LSA) provide data that can be used to evaluate outcomes of policies and practices. When it comes to making statistical inferences using LSA data, new developments in Bayesian modeling and Bayesian prediction techniques are needed to fully exploit the data information, while accounting for complex dependency structures. For LSA studies with a pretest-posttest design, an innovative Bayesian multivariate multilevel IRT modeling framework is proposed to compare groups, measure change, and identify effects of relevant background variables. This Bayesian methodology has been specifically developed for a comprehensive, two-year, schoolwide intervention to improve data-based decision-making (DBDM). Two unique applications are discussed using the newly developed Bayesian modeling methods. When the main goal is prediction, the focus on a particular theoretical model is of less importance than establishing predictive accuracy. It is shown that Bayesian model averaging (BMA) techniques outperform any other model when it comes to making predictions. The BMA techniques is illustrated using PISA data. This session provides an overview of recent developments in Bayesian modeling and prediction methods using LSA data. The combined set of techniques provide innovative ways to analyse LSA data, while addressing complex features associated with LSA studies.

Bayesian Multivariate Multilevel Models for Pretest-Posttest LSA Data
Jean-Paul Fox, University of Twente

The Development of Teaching Quality during DBDM Intervention: Pretest-Posttest Multilevel IRT Approach
Trynke Keuning, University of Twente; Marieke van Geel, University of Twente; Adri Visscher, University of Twente; Jean-Paul Fox, University of Twente

The Impact of a Data-Based Decision Making Intervention on Educators' Data Literacy
Marieke van Geel, University of Twente; Trynke Keuning, University of Twente; Adri Visscher, University of Twente; Jean-Paul Fox, University of Twente

Bayesian Model Averaging for Building Predictive Models with Large-Scale Educational Assessments
David Kaplan, University of Wisconsin-Madison

Friday, April 28, 2017
12:25 PM–1:55 PM, Conference Room 3&4, Meeting Room Level, Coordinated Session, C3

Psychometric Issues In Multidimensional IRT

Session Organizer: Ming Lei, American Institutes for Research
Session Chair: Okan Bulut, University of Alberta
Session Discussant: Mark Reckase, Michigan State University

The purpose of this symposium is to address psychometric issues in multidimensional IRT (MIRT) using real data from operational testing. Four papers are included. The first paper compares methods (logistic regression procedure, generalized multi-group bifactor model, and the IRT-based internal measurement) in multidimensional DIF detection. The second paper shows the known paradoxical situation that an additional correct response can lower ability estimate while an incorrect response can increase ability estimate is an example of explaining-away phenomenon, which in MIRT, can be related to general model properties. The third paper assesses the performance of posterior predictive model checking (PPMC) method in MIRT. The fourth paper proposed a procedure to replace scaling transformation matrices with scale constants. The English Language Proficiency Assessment for the 21 Century (ELPA21) is used in all of the studies. ELPA21 includes four dimensions, reading, writing, listening, and speaking. MIRT (3PL for dichotomous items and GR model for graded response items) is used item parameter calibration. Bifactor model is used for the overall ability estimation.

Detecting Multidimensional Differential Item Functioning
Ming Lei, American Institutes for Research; Hyesuk Jang, American Institutes for Research; Okan Bulut, University of Alberta

Explaining Away and Paradoxical Results
Frank Rijmen, American Institutes for Research; Peter van Rijn, Educational Testing Service

Posterior Predictive Model Checking in Detecting Testlet Clustering Effects
Chao Xie, American Institutes for Research; Tongyun Li, Educational Testing Service

Developing Scalar Linking Constants in Multidimensional IRT Modeling
Dandan Liao, University of Maryland; Hong Jiao, University of Maryland; Ming Lei, American Institutes for Research

Friday, April 28, 2017
12:25 PM–1:55 PM, Salon B, Meeting Room Level, Coordinated Session, C4

Engineered Cut Scores: Aligning Standard Setting Methodology with Contemporary Assessment Design Principles

Session Chair: Daniel Lewis, Pacific Metrics Corporation
Session Discussant: Gregory Cizek, University of North Carolina

Evidence Centered Design (ECD) and Assessment Engineering (AE) are contemporary models that provide principled frameworks for designing and developing assessments. The presenters in this session will describe a new approach to standard setting—Engineered Cut Scores—that is well-aligned with the principles of ECD and AE. Engineered Cut Scores supports the estimation of cut scores directly as an outcome of an enhanced ALD and item writing process. That is, if standard setting methodology were fully aligned with contemporary principled assessment design, a separate standard setting workshop may not be needed. In this session, the presenters (a) provide the historical context for current standard setting methods, (b) provide a strong rationale for the embedding of standard setting in the test development cycle under ECD and AE, (c) describe the methodological details of such an embedded standard setting process—Engineered Cut Scores, (d) discuss the systematic approaches and methods that would be required to support achievement level descriptor and item writing to support ECS, and (e) discuss policy considerations that a sponsoring agency should consider under this new method. Aligning modern standard setting with contemporary test design, as proposed here, is expected to enhance the validity of states' assessment programs.

The Historical and Methodological Context for Engineered Cut Scores
Daniel Lewis, Pacific Metrics Corporation

Aligning Knowledge and Skill Requirements in ALDs and Item Response Demands
Steve Ferrara, Measured Progress

Policy Implications of Engineered Cut Scores
Wesley Bruce, Independent Educational Consultant

Friday, April 28, 2017
12:25 PM–1:55 PM, Salon C, Meeting Room Level, Coordinated Session, C5

Building validity arguments for domestic and international college learning outcomes assessments

Session Chair: Guangming Ling, Educational Testing Service

The last two decades have witnessed a wide adoption of standardized assessment tools to assess learning outcomes and facilitate improvement at higher education institutions in the United States and worldwide. The assessment tools either focus on generic college-level skills, such as the ETS HEIghten™ suite of assessments that measure cognitive and noncognitive skills, or target domain-specific knowledge and skills, such as the ETS Major Field Tests. A critical step in building a validity argument for such uses is to collect empirical evidence to support the claim that these assessments capture student learning in the targeted domains as expected, regardless of the skills and competencies being generic or domain-specific, in either domestic or international contexts. For this purpose, this coordinated session aims to first provide an overview of the current uses of generic and domain-specific outcomes assessments and related challenges, followed by presentations of four empirical validation studies of these assessments, for both two- and four-year college students, in both domestic and international contexts.

An overview of college students' learning outcomes assessment and some challenges
Guangming Ling, Educational Testing Service; Katrina Roohr, Educational Testing Service; Ou Lydia Liu, Educational Testing Service

Evaluating validity evidence for a learning outcomes assessment at community colleges
Katrina Roohr, Educational Testing Service; Kri Burkander, Educational Testing Service

Validating the use of HEIghten™ Quantitative Literacy assessment in China
James Mason, University of California at BerkEley; Lin Gu, Educational Testing Service; Ou Lydia Liu, Educational Testing Service; Amy Shaw, Rice University; Prashant Loyalka, Stanford University

Building a validity argument for the Major Field Test in Biology
Francis Rick, University of Massachusetts, Amherst; Guangming Ling, Educational Testing Service; Zhen Wang, Educational Testing Service

Linking MFT Business scores across forms in domestic and international samples
Zhen Wang, Educational Testing Service; Guangming Ling, Educational Testing Service; Francis Rick, University of Massachusetts, Amherst

Friday, April 28, 2017
12:25 PM–1:55 PM, Salon D, Meeting Room Level, Paper Session, C6

Issues in Model-data (mis)fit

Cautions on Using Model Fit to Choose Number of Factors in EFA

Amanda Montoya, The Ohio State University; Michael Edwards, The Ohio State University

Researchers often use fit indices to determine the number of factors to retain in EFA. We caution against this, as existing cut-off recommendations do not work well and the fit statistics are dependent on attributes unrelated to the number of factors (e.g., overall misfit of the model and correlated residuals).

Information Criteria for Item Response Theory Model Selection

Seock-Ho Kim, The University of Georgia

A review of various indices of information criteria is presented for IRT model selection. Using example data it explicates why IRT computer programs report different values of the likelihood functions and consequently different indices of information criteria. Uses of information criteria for more general and complicated IRT models are discussed.

The Impact of Misfitting Responses on Large Scale Assessments

Ying Cui, University of Alberta; Xin Tao, Beijing Normal University; Ping Chen, Beijing Normal University

The goal of this study was to investigate the impact of the presence of misfitting responses on tests using data from a national assessment program in China. Results from this study shed lights on the usefulness of person fit analysis in validating test results in large scale assessment settings.

Using Limited Information Fit Statistics/Indices for Evaluation of Measurement Invariance

Mengyao Cui, Florida State University; Yanyun Yang, Florida State University

This study investigates the performances of the limited information fit statistic (M2) and its corresponding fit indices in measurement invariance evaluation within multiple-group IRT framework. A simulation study examines sampling distributions of M2 and its descriptive fit indices, and their sensitivities to lack of measurement invariance under various conditions.

Friday, April 28, 2017
12:25 PM–1:55 PM, Salon K, Meeting Room Level, Paper Session, C7

Modeling non-cognitive traits with IRT

A Hierarchical Domain Framework of Test Validity and Validation
Jinsong Chen, SYSU

The framework offers flexibility and scalability in validating score meaning and use. It provides (a) a unique meaning domain for test scores based on the relative nature of attributes, (b) differentiation of score meaning and score use, and (c) a general framework with domains to signify the degree of validity.

Ability Estimates for Multi-Unidimensional Pairwise-Preference Model for Personality Assessment
Lihua Yao, Defense ManPower Data Center; Rich Riemer, Defense Manpower Data Center; Daniel Segall, Defense Manpower Data Center; Mary Pommerich, Defense Manpower Data Center

Multi-Unidimensional Pairwise-Preference(MUPP) Model has been used for Personality Assessment. In this study, examinees ability parameters are estimated by different methods and their precision are compared. Both real data and simulated data are applied. For the simulation, the test length, test structure, and test dimensions are varied.

Dimensionality and Item Hierarchy of the 12-Item Theory of Mind Scale
Beyza Aksu Dunya, University of Illinois at Chicago; Everett Smith, University of Illinois at Chicago; Clark McKown, Rush University

The purpose of this paper is to investigate the dimensionality and item difficulty hierarchy of the 12-item Theory of Mind (ToM) scale. The ToM scale was developed for measuring progression of conceptual achievements that mark social cognitive understanding in normally developing children for grades 1 to 4.

Impact of Aberrant Responding and Its Detection in Forced-Choice Noncognitive Assessments
Sooyeon Kim, Educational Testing Service; Tim Moses, College Board

The purpose of this study is to assess the impact of aberrant responses on the estimation accuracy, as a function of aberrance types and proportions, in forced-choice personality measures. Three person-fit indices are compared in terms of their effectiveness of detecting examinees' aberrant behaviors when various aberrant responses are present.

RESET: A Special Education Evaluation Tool that Improves Instruction Through Feedback
Evelyn Johnson, Boise State University; Angela Crawford, Boise State University; Laura Moylan, Boise State University

Recognizing Effective Special Education Teachers (RESET) consists of observation rubrics designed to provide special education teachers with feedback and an evaluative score on their implementation of EBPs. This presentation provides an overview of RESET, presents findings on its psychometric properties, and the effect of feedback on teacher practice.

Friday, April 28, 2017
12:25 PM–1:55 PM, Salon L, Meeting Room Level, Paper Session, C8

Bayesian Modeling

A Binomial-lognormal Latent Model to Measure Reading Speed

Akihito Kamata, Southern Methodist University; Yusuf Kara, Anadolu University; Cornelis Potgieter, Southern Methodist University

This paper proposes and evaluates a new psychometric model to measure reading speed by a binomial-lognormal latent variable model adopting the Bayesian Approach. Analysis of simulated data set indicated that the model and its parameter estimation are promising and warrant further investigations for practical use.

Bayesian SEM Analysis by Using Polya-Gamma Method

Seohyun Kim, The University of Georgia; Zhenqiu (Laura) Lu, The University of Georgia

The Polya-Gamma (PG) estimation method was recently proposed (Polson et al., 2013) to improve the simplicity and efficiency of the traditional Bayesian sampling method for statistical models with dichotomous outcomes. In this paper, we extended the PG estimation method to SEM with a focus on investigating the effective sample sizes.

Investigating the Power of Bayesian Multiplicity Corrections with Heterogeneous Group Variances

Weldon Smith, University of Nebraska-Lincoln; Michael Zweifel, University of Nebraska-Lincoln

Multiple comparison procedures control Type I error inflation, but also significantly reduce power. Variance heterogeneity among groups negatively affects the procedures. The current study investigated how Bayesian multilevel models, which can model heterogeneity, compare to traditional procedures in terms of Type I error rates and power across varied conditions.

On the implications of using truncated priors in Bayesian inference

Xiang Liu, Teachers College, Columbia University; Zhou Zhou, Teachers College, Columbia University; Hui Soo Chae, Teachers College, Columbia University; Gary Natriello, Teachers College, Columbia University

Parameter constraints are often necessary to identify psychometric models. The constraints are commonly imposed through truncating priors. However, this will sometimes result in unintended consequences to posterior sampling. We propose a solution based on post processing using the samples from the unconstrained model. An unfolding model is used to illustrate.

Friday, April 28, 2017
12:25 PM–1:55 PM, Salon M, Meeting Room Level, Paper Session, C9

Automatic Item Generation

Session Chair: Tim Konold, University of Virginia

A Multilevel MT-MM Approach for Estimating Contextual Influences on Informant Effects
Tim Konold, University of Virginia

An approach for extracting common method variance from structurally different informant ratings is presented with an extension for estimating the degree to which they are associated with contextual factors. Results are based on structurally different and interchangeable students (N = 45,641) and teachers (N = 12,808) from 302 schools.

Bayesian Expectation-Maximization-Maximization for the 3PLM
Shaoyang Guo, Jiangxi Normal University; Chanjin Zheng, Jiangxi Normal University; Xiaohong Gao, ACT Inc.; Hua-Hua Chang, University of Illinois at Urbana-Champaign

Expectation-Maximization-Maximization (EMM) can produce satisfactory estimates for 3PLM, but we observed some estimates still blow up. This study aims to take advantage of both the EMM and the Bayesian EM(BEM) and proposes the Bayesian EMM(BEMM). The simulation indicates that BEMM has a better performance than BEM and EMM.

Evaluating Text Similarity of Generated Items Using Graph
Xinxin Zhang, University of Alberta; Mark Gierl, University of Alberta; Marie Wiberg, Umeå School of Business and Economics

This research verifies that the complexity of a graph, which is a visual representation of the model used to generate items, reflects the text similarity of generated items, thus justifying the use of graph to evaluate text similarity of generated items as an innovative method.

IRT in SPSS Using the SPIRIT Macro
Jack DiTrapani, The Ohio State University; Nicholas Rockwood, The Ohio State University; Hyeon-Joo Oh, Educational Testing Service

We present a SPSS macro, called SPIRIT, for flexible and powerful item response analysis in SPSS. SPIRIT allows researchers to fit various one-parameter family models that can be created with multiple dimensions, multiple groups, person covariates, DIF parameters, item predictors (and error term), etc. for binary and polytomous item responses.

Rationale Generation: An Expansion of the Item Generation Framework
Mark Gierl, University of Alberta; Hollis Lai, University of Alberta

This study expands the automatic item generation framework to address the content-development challenge of creating the rationale for each generated item required for computerized formative testing. We describe and demonstrate a new method for producing the rationales for the newly generated surgical education test items.

Friday, April 28, 2017
12:25 PM–1:55 PM, Salon J, Meeting Room Level, Electronic Board Session, Paper Session, C10

Electronic Board #2

Mining Students' Constructed Response Answers

Minho Kwak, University of Georgia; Allan Cohen, University of Georgia; Seohyun Kim, University of Georgia

Latent Dirichlet allocation (LDA) is used to analyze middle grades students' constructed response answers on a test of science inquiry knowledge. Initial results indicate that LDA can provide useful information about the effects of instruction. Students' post-test responses showed an increase in use of words that aligned closely with instruction.

Electronic Board #3

Standard Error estimation using Bootstrap approaches for cognitive diagnosis models

Wenjing Guo, City University of New York; Wenchao Ma, Rutgers, The State University of New Jersey; Jimmy de la Torre, The University of Hong Kong

Existing procedures for estimating standard errors of item parameters for cognitive diagnosis models have some limitations. This study evaluates the performance of bootstrap resampling approach in estimating standard errors under various conditions. Both simulated and real data are considered in this study.

Electronic Board #4

Anchor-Based Method for Judgmentally Estimating Item Difficulty Parameters Using a Q-Sort Procedure

Sharon Frey, Houghton Mifflin Harcourt; Joni Lakin, Auburn University; Stephen Murphy, Houghton Mifflin Harcourt; Emmett Cartwright, Houghton Mifflin Harcourt

The study demonstrates the use of Q-Sort methodology to collect expert judgments of item difficulty without field item tryouts. The expert ratings and known item difficulties of embedded anchors are used to estimate the item difficulties for non-anchors in order to inform decisions for the development of new test forms.

Electronic Board #6

Selecting bandwidths in kernel equating with different tests and test distributions

Gabriel Wallin, Umeå School of Business and Economics, Umeå University; Jenny Häggström, Umeå School of Business and Economics; Marie Wiberg, Umeå School of Business and Economics

In kernel equating one needs to select the bandwidth of the kernel smoother when continuizing the cumulative distribution functions of the test scores. In this study we have evaluated and compared the existing methods for selecting the bandwidth, and furthermore proposed a new way of selecting it.

Electronic Board #7

Mode Comparability Study of Second Graders across Four Subject Areas

XIANGDONG LIU, THE UNIVERISTY OF IOWA; CATHERINE WELCH, THE UNIVERSITY OF IOWA; STEPHEN DUNBAR, THE UNIVERSITY OF IOWA

Mode comparability study of computerized and paper-and-pencil versions of a test was conducted among four subject areas at the second grade level. The results indicate that CBTs were generally easier than PPTs on language, mathematics and science tests. The study also explores whether item presentation characteristics affect mode differences.

Electronic Board #8

Paul Jewsbury, Educational Testing Service; Peter van Rijn, Educational Testing Service

Paul Jewsbury, Educational Testing Service; Peter van Rijn, Educational Testing Service; Akihito Kamata, Southern Methodist University

Simple structured tests may be modelled with a series of unidimensional IRT models or with a single multidimensional IRT (MIRT) model. The MIRT approach was found to be more effective for a Multistage Testing (MST) design in estimating all item parameters on a shared metric.

Electronic Board #9

Evaluating Scale Stability in IRT Complex Linkage Plans with Item Parameter Drift

Christiana Akande, University of Florida; David Miller, University of Florida; Anne Manley, University of Florida

This study investigates the degree to which anchor length, amount of IPD and length of equating links simultaneously and independently affect the stability of equating results. A NEAT design was utilized. Results will assist practitioners in making optimal decisions about anchor length, and about handling IPD and multiple link concerns.

Electronic Board #10

Using Response Time Models to Identify Possible Classroom Level Cheating

Jessalyn Smith, DRC; Keith Boughton, DRC

Test security breaches compromise the psychometric integrity of testing programs and there is value in finding methods to detect and prevent breaches. The purpose of this study is to add to the literature by using response time (RT) modeling in the context of identifying classroom/school level test security breaches.

Electronic Board #11

Talking to Test Takers – What Can Test Developers Learn?

Matthew Schultz, American Institute of Certified Public Accountants; Jacqueline Leighton, University of Alberta; Joshua Stopek, American Institute of Certified Public Accountants

Testing organizations can gather informative feedback from an under-utilized segment of experts – current and prospective test takers. To that end, a framework of test taker/testing organization communication intersections is proposed, highlighting the goals of each, information gathered, and how the results inform ongoing development, primarily focused on test validity.

Electronic Board #12

Developing Parallel Paper Forms for Online-Focused Assessments

Jennifer Beimers, Pearson; Jasmine Carey, Colorado Department of Education; Joyce Zurkowski, Colorado Department of Education

There has been a shift in large scale testing from paper-and-pencil assessments to an online-focus. But regardless of how online-focused an assessment is, a paper-based equivalent form is often needed. The purpose of this study is to describe a state testing program's strategy and to investigate stability across modes.

Electronic Board #13

Comparing Ability Estimation Precision of Routing Methods for Multistage Adaptive Testing

Evgeniya Reshetnyak, Fordham University; Xinhui Xiong, AICPA

The purpose of the current study is to evaluate and compare the precision of the ability estimation of six routing strategies (IRT, CART, FI, K-L, DPI and Bayesian) in the context of adaptive MST design using the data from a licensure exam and simulations.

Electronic Board #14

Eliminating Item Parameter Drift of Item Position Change Through Field Test Design

Anna Topczewski, GED Testing Service

Item parameter estimates are variant, in part, due to item position change. To constrain item parameter drift due to position change, considerable constraints are put on the test construction process. This study evaluates item parameter drift under a field test conditions designed to eliminate the effect of item position change.

Electronic Board #15

Two item-matching heuristic methods to assemble multiple parallel forms with common items

Pei-Hua Chen, National Chiao Tung University; Cheng-Yi Huang, National Chiao Tung University

How to handle common items in heuristics methods for multiple forms assembly has not been discussed in the test assembly literature. The purpose of this study is to propose two item matching heuristic methods to deal with two kinds of common item requirements in assembling multiple parallel forms.

Electronic Board #16

Modeling Item-Position Effects in Large-Scale Assessment Using the Multidimensional IRT Model

Xugang Nie, Beijing Normal University; Ping Chen, Beijing Normal University

The objective of this study is to use the MRMLC to model and analyze the item-position effects in large-scale assessments. According to Embretson (1991)'s design in parameter estimation, we designed two groups of subjects and compared two methods for estimating the ability parameters in MRMLC.

Electronic Board #17

Using Rater Judgements to Estimate Item Difficulty of Cognitive Tasks

Emmett Cartwright, Houghton Mifflin Harcourt; Stephen Murphy, Houghton Mifflin Harcourt; Joni Lakin, Auburn University; Sharon Frey, Houghton Mifflin Harcourt

This study used a Q-Sort methodology to collect rater judgments of both item difficulty and the item characteristics those underlie estimations of perceived difficulty. The findings contribute to research on the use of alternative methods of data collection and the analysis of item characteristics in test design.

Electronic Board #18

Evaluating Item Parameter Drift Methods in a CAT Environment

Kevin Cappaert, Pearson; Yao Wen, University of Wisconsin - Milwaukee; Yu-Feng Chang, Minnesota Department of Education

A lingering issue in computer adaptive testing (CAT) is how to best evaluate item parameter drift (IPD) given the narrow student ability distribution of students assessed by a given item. The present study evaluates four sets of quadrature weights for the D2 method of IPD detection.

Electronic Board #19

Difficulty of the sentence equivalence item type: Implications for a validity argument

Isaac Bejar, ETS; Paul Deane, ETS; Michael Flor, ETS; Jing Chen, HumRRO

The report is the first systematic evaluation of the sentence equivalence item type introduced by the revised GRE®. We adopt a validity framework to guide our investigation based on Kane's (2006). We provide evidence of the construct representation of the sentence equivalence item type.

Electronic Board #20

Generalized Scripted Testing—One System for All

Chunxin Wang, ACT Inc.; Jie Li, ACT Inc.; Wugen Dai, ACT Inc; Nancy Petersen, ACT Inc.; Lisa Gawlick, ACT Inc.

Generalized Scripted Testing (GST) is a comprehensive test assembly method and system. This study substantiates the versatility and flexibility of the GST system by providing applications and examples of GST for linear, adaptive and multistage testing as well as for aptitude, diagnostic and licensure exams.

Electronic Board #21

Exploratory Evaluation of Online Accessibility Tools for Magnification, Color Contrast, and Text-to-Speech

Anne Davidson, Anne H. Davidson; Nathan Wall, eMetric

The study purpose is to explore how students used accessibility tools of magnification, color contrast, and text-to-speech during online testing. Descriptive analysis of data from one state's middle school population (≈135,000) will improve understanding of how students use accessibility features to assist with the specification of accessibility features.

Electronic Board #22

Technology-Enhanced Items: Validity of Inferences and Measurement of Fifth-Grade Geometry

Jessica Masters, Measured Progress

Technology-enhanced items have potential to provide improved measurement of constructs. More research is needed to evaluate whether these items lead to valid inferences and provide improved measurement over traditional items. This paper explores these questions in the context of fifth-grade geometry using quantitative analysis (including classical test theory and IRT).

Friday, April 28, 2017
2:15 PM–3:45 PM, Salon A, Meeting Room Level, Coordinated Session, D1

Contemporary Issues with Interim Assessments

Session Chair: Thanos Patelis, Center for Assessment
Session Discussant: Joan Herman, UCLA/CRESST

Interim assessments are a broad group of assessments that have a few characteristics in common: they are given between summative assessments, usually more than once to the same students, and the results are intended to be aggregated across students (Perie et al., 2009). Although state summative assessments have long received much attention, according to some estimates, many more interim than summative assessments are administered each year. This symposium addresses contemporary issues within interim assessments: What are the various aspects of ESSA crossed with the critical evidence required for peer review? How can they be evaluated when their uses and administration conditions vary significantly? What are essential features of assessments used for various purposes, specifically for instruction, to measure growth, and to predict performance on a future assessment?

Flexibility and Innovation with Evidence Considerations: Design in the Era of ESSA
Karen Barton, NWEA

Evaluating Interim Assessments: Non-standardization, Quality, and Usefulness
Brian Gong, Center for Assessment

Essential Features of Interim Assessments used for Various Purposes
Marianne Perie, University of Kansas

Friday, April 28, 2017
2:15 PM–3:45 PM, Conference Room 1&2, Meeting Room Level, Coordinated Session, D2

Classroom realities that need to be understood to make good assessments

Session Chair: Gavin Brown, The University of Auckland

The ideal theoretical model of human behavior in formal testing situations is that each test-taker has had high quality instruction and preparation, is highly motivated, pays 100% attention throughout the test event, and has had sufficient personal environmental support (i.e., sleep, nutrition, etc.) to produce best possible results. These assumptions are generally valid in high-stakes formal certification examinations, but much less so in low-stakes testing and the kinds of assessment carried out in real-world classroom settings. Assessment research in school, classroom, and higher education settings shows that factors within (a) the individual psychology of teachers and students and (b) the social psychology of learning environments affect the validity of scores, interpretations, and decisions. This symposium provides overviews of recent research about the impact of 'in vivo' contexts on assessment processes, practices, and policies. This is especially important as policies that advocate low-stakes and formative assessments are introduced and implemented across all levels of education.

Enhancing Assessment Capability of Teachers
Helen Timperley, The University of Auckland; Judy Parr, The University of Auckland

Student Perceptions of Assessment
James McMillan, Virginia Commonwealth University

Assessments that Work in Classrooms
Susan Brookhart, Brookhart Enterprises LLC

Bidirectional Relationships between Assessment and Sociocultural Context
Kadriye Ercikan, University of British Columbia; Guillermo Solano-Flores, Stanford University

Reflections on the future of assessment as a human and social endeavour
Gavin Brown, The University of Auckland; Lois Harris, Central Queensland University

Friday, April 28, 2017
2:15 PM–3:45 PM, Conference Room 3&4, Meeting Room Level, Coordinated Session, D3

Multilevel Latent Variable Modeling Strategies for Handling Error in Predictors

Session Chair: Kilchan Choi, CRESST/UCLA
Session Discussant: Daniel McCaffrey, ETS

Measurement error can weaken the validity of inferences from student assessment data, reduce the statistical power of impact studies, and diminish the ability of researchers to identify causal mechanisms that lead to an intervention improving the desired outcome. We propose a general statistical framework based on multilevel latent variable modeling (MLVM) to address measurement error related issues in educational assessment and evaluation studies. The first paper details how MLVM strategies can help overcome such problems and discusses the care that must be exercised both from a design and analysis standpoint in implementing such strategies. The second paper proposes a multilevel item factor model with multiple latent predictors in the level-1 model. The third paper focuses on the sensitivity of results to priors for the three-level latent variable model with the level-two measurement model in a Bayesian MCMC estimation. A small-scale targeted simulation examines the degrees of bias and rates of coverage of parameters of interest by comparing the results from employing non-informative priors vs. mildly-informative priors for key variance components and fixed effects. The last paper proposes a multilevel latent variable mediation model in which a latent mediator is modeled as a function of treatment in multisite cluster randomized trials.

Multilevel/MCMC-based Strategies for Handling Error in Predictors: Pressure Points and Implementation Recommendations
Michael Seltzer, UCLA; Ji Seung Yang, University of Maryland; Kilchan Choi, CRESST/UCLA

Multilevel Item Factor Model with Multiple Latent Predictors in Level-1 Model
Li Cai, CRESST/UCLA; Kilchan Choi, CRESST/UCLA

Sensitivity of Results to Priors: Three-Level Latent Variable Model with Level-Two-Measurement Model
Kilchan Choi, CRESST/UCLA; Michael Seltzer, UCLA

Treating Mediator as a Latent Variable in Multisite Cluster Randomized Trials
Kilchan Choi, CRESST/UCLA; Michael Seltzer, UCLA; Li Cai, CRESST/UCLA

Friday, April 28, 2017
2:15 PM–3:45 PM, Salon B, Meeting Room Level, Coordinated Session, D4

Improving Technical Quality in K-12 Computerized-Adaptive Tests

Session Chair: Liru Zhang, Delaware Department of Education
Session Discussant: Denny Way, Pearson

Computerized Adaptive Testing (CAT) has been increasingly adopted in large-scale assessments because of its advantages of high efficiency and great precision in measurement especially for low-achieving and high-achieving students. When individual tests are assembled simultaneously to a large population, however, the optimization becomes a big challenge in operation. Many special features in K-12 education, for example, broader content standards, within-grade curriculum, diversity of population, and wide range of academic achievement, may add additional barriers to the implementations of CAT. In this session, four researchers will discuss technical issues in K-12 CAT on item pool design to measure the full-range of the Common Core State Standards with balance in content and cognitive complexity; the influences over measurement precision and student performance when alternative IRT models are applied separately in item selection and in scoring; the appropriateness of the LOSS/HOSS that are based on standalone, fixed-form field tests data for the operational adaptive tests; and the comparability of test scores in the transition from an paper/pencil test to an online multi-stage adaptive test for English language learners. A highly regarded discussant in this field will discuss key technical considerations in the K-12 adaptive testing.

Item Pool Design for Large-Scale Computerized-Adaptive Tests
Marty McCall, Smarter Balanced Assessment Consortium

Effects of Ignoring Discrimination Parameter in CAT Item Selection on Student Scores
Shudong Wang, NWEA

Considerations in Setting LOSS/HOSS for Computerized-Adaptive Tests
Liru Zhang, Delaware Department of Education

Effects of Ignoring Discrimination Parameter in CAT Item Selection on Student Scores
Liru Zhang, Delaware Department of Education

Considerations in Setting LOSS/HOSS for Computerized-Adaptive Tests
Shudong Wang, NWEA

Transition from Paper to Adaptive Testing – An ELP Assessment Experience
Gary Cook, Wisconsin Center for Education Research – WIDA Consortium; Denny Way, Pearson; Liru Zhang, Delaware Department of Education; Kyoungwon Bishop, Wisconsin Center for Education Research – WIDA Consortium

Friday, April 28, 2017
2:15 PM–3:45 PM, Salon C, Meeting Room Level, Coordinated Session, D5

Reworking the Multiple-choice Item to Improve, Teaching, Learning and Program Evaluation

Session Organizer: Mark Davison, University of Minnesota
Session Chair: Ben Seipel, University of Wisconsin River Falls
Session Discussant: Steve Culpepper, University of Illinois

A new, computer administered reading comprehension test, the Multiple-choice Online Causal Comprehension Assessment (MOCCA) uses an innovative multiple-choice reading item format. The format is innovative in four respects. (1) Instead of two types of item responses, correct and incorrect, it has three item response types, correct and two types of incorrect responses. (2) It has only one item per reading passage. (3) It uses a cloze format in which a sentence is missing from each reading passage, and the student must select the sentence that best completes the passage. (4) Item response times are recorded. The first paper describes these innovations. The second and third papers present data on the interpretation, fit, and measurement precision (information functions) for item response models designed to provide reading comprehension scores plus diagnostic scores about types of errors made by poor comprehenders and the comprehension efficiency of good comprehenders. The item response models pose a new type of missing data problem. The last paper describes the missing data problem, and presents simulation data on the recovery of person parameters in the presence of such missing data.

The Innovative Item Format of the Multiple-choice Online Causal Comprehension Assessment
Sarah Carlson, University of Oregon; Ben Seipel, University of Wisconsin River Falls; Gina Biancarosa, University of Oregon

Item Response Modeling of Incorrect Answer Choices
Gina Biancarosa, University of Oregon; Bowen Liu, University of Minnesota; Ben Seipel, University of Wisconsin River Falls; Sarah Carlson, University of Oregon

Item Response Models Incorporating Item Response Times
Mark Davison, University of Minnesota; Qinjun Wang, University of Minnesota; Shiyang Su, University of Minnesota

A Simulation Study of MOCCA's Response Conditional IRT Models
Qinjun Wang, University of Minnesota; Bowen Liu, University of Minnesota; Shiyang Su, University of Minnesota; Mark Davison, University of Minnesota

Friday, April 28, 2017
2:15 PM–3:45 PM, Salon D, Meeting Room Level, Paper Session, D6

Dealing with Missing Data

Session Chair: Quinn Lathrop, Advanced Computing and Data Science Lab Pearson

Alternative Multiple Imputation Inference for Categorical Structural Equation Modeling

Seungwon Chung, University of California, Los Angeles; Li Cai, UCLA/CRESST

This study proposes an alternative approach to missing data in structural equation modeling. It extends work by Lee and Cai (2012) on multiple imputation for continuous variables by incorporating a method to account for ordered categorical variables, which we frequently encounter in survey research.

Imputing 12th-Grade NAEP Mathematics Scores for the Full HSLS Sample

Burhan Ogut, American Institutes for Research; George Bohrnstedt, American Institutes for Research; Markus Broer, American Institutes for Research

In 2013, about 17% of the students participating in the High School Longitudinal Study of 2009 also took the National Assessment of Educational Progress grade 12 mathematics assessment. The purpose of this study is to determine the feasibility to impute NAEP mathematics scores for the remainder of the HSLS sample.

Making Psychometric Inferences with SVD when Data are Missing Not at Random

Quinn Lathrop, Advanced Computing and Data Science Lab Pearson

Singular Value Decomposition (SVD), a technique popular in data mining, can make psychometric inferences that are generally robust to missing data (including MNAR). Analytical results provide a psychometric interpretation of SVD as person ability and item easiness ordinal estimators. Simulations show that SVD can outperform both IRT- and CTT-based inferences.

Nested Multiple Imputation Procedures for Estimating Variance Contributions from IRT Model Uncertainty

Lauren Harrell, National Center for Education Statistics

In large-scale survey assessments, plausible values are drawn assuming fixed IRT parameter estimates. Nested multiple imputation procedures are modified to evaluate and adjust for uncertainty in IRT parameter estimation and impacts on variance of population quantity estimates. Simulations of IRT parameter uncertainty conditions and applications to linking error are discussed.

Trait Estimation with Imputation When Data Are Missing Not At Random

Rose Stafford, The University of Texas at Austin; Christopher Runyon, The University of Texas at Austin; Jodi Casabianca, The University of Texas at Austin; Barbara Dodd, The University of Texas at Austin

This research investigates trait level estimation under the rating scale model when data are missing not at random. Three imputation methods are investigated: (a) multiple imputation, (b) nearest-neighbor hot deck imputation, and (c) multiple hot deck imputation. We compare results across three levels of missingness crossed with three scale lengths.

Friday, April 28, 2017
2:15 PM–3:45 PM, Salon K, Meeting Room Level, Paper Session, D7

Issues in Standard Setting 1

Session Chair: Michael Bunch, Measurement Incorporated

A Graphical Alternative to Traditional Borderline Examinee Descriptors for Licensure Tests

Priya Kannan, Educational Testing Service; Richard Tannenbaum, Educational Testing Service; Delano Hebert, Educational Testing Service

We evaluated an alternative to traditional borderline examinee descriptors, called Anchored Graphical Representations (AGRs; Authors, 2015). AGRs provide both text and visual organizers by each tested domain to contextualize the meaning of the borderline examinee descriptors. Results from two studies support the value of the AGRs.

An Experimental Study of the Internal Consistency of Bookmark Standard Setting Judgments

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

In bookmark standard setting, the difficulty of ordered item booklets should not affect the resulting cutscores. Yet, in an experiment in which judges were randomly assigned to easy or difficult booklets, booklet difficulty did systematically affect judgments. This finding calls into question prevalent beliefs about the meaning of these judgments.

Establishing Cut Scores via Latent Class Analysis for a Large Scale Assessment

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

This study employs latent class analysis for setting performance standards for a large scale science assessment and investigates whether the conclusions are comparable to those from previously established standards via modified-Angoff Method. The findings establish evidence supporting the results of the former standard setting process.

Standard Setting to Go: Online Body of Work with Articulation

Michael Bunch, Measurement Incorporated

This paper describes a standard setting activity for a series of writing assessments for grades 3-10, conducted entirely online. Software development and deployment, management of training webinars and inter-round discussions, a vertical articulation webinar, and final evaluation of the process by panelists, are described in detail.

Validity Evidence for the Performance Profile Standard Setting Method

Lei Wan, College Board; Luz Bay, College Board; Deanna Morgan, College Board

The performance profile method (PPM) is a relatively new standard-setting method. The validity evidence supporting this method was scarcely reported. This paper describes the method, and demonstrates the validity of this method based on procedural, internal, and external sources of validity evidence (Kane, 1994, 2001).

Friday, April 28, 2017
2:15 PM–3:45 PM, Salon L, Meeting Room Level, Paper Session, D8

Validity Issues in Test Design

Session Chair: Okan Bulut, University of Alberta

A Generalized Approach to Measuring Test-Taking Effort on Computer-Based Tests

Steven Wise, Northwest Evaluation Association

When CBTs are used, test taker behavior can be used to identify item responses that the test taker responded to in a disengaged fashion. This study extends previous research by proposing a generalized approach to identifying non-effort that can be used with multiple-choice items, constructed responses, and omitted answers.

Investigating Side-Effects of CBA in LSA: Comparison of Test-Taking Engagement between Modes

Ulf Kroehne, German Institute for International Educational Research (DIPF); Frank Goldhammer, German Institute for International Educational Research (DIPF)

Rapid guessing behavior is compared between modes (computer vs. paper) and settings (proctored vs. un-proctored) for two test-domains (Science and ICT Literacy). The analysis of individual response-times reveals a mode-effect on engagement for Science, a general mode-effect of proctoring for both domains and an additional gender effect for un-proctored testing.

Issues for Developing Science Contextualized Items

Min Li, University of Washington at Seattle; Maria Ruiz-Primo, Stanford University; Dongsheng Dong, University of Washington at Seattle; Jim Minstrell, Facet Innovations Inc.; Xiaoming Zhai, Beijing Normal University; Phonraphee Thummaphan, University of Washington

Despite the widespread use of contexts in science testing, their utility, practice, and underlying assumptions have been called into question. In this paper, we propose a conceptual framework for evaluating the quality of contextualized items and provide empirical evidence by analyzing a pool of released items based on expert reviews.

The Achilles' Heel of Multiple-Choice Items: Distractors

Okan Bulut, University of Alberta; Mark Gierl, University of Alberta; Qi Guo, University of Alberta; Xinxin Zhang, University of Alberta

Multiple-choice (MC) items have been commonly used in both classroom assessments and large-scale standardized tests for many years. Yet, there is little consensus on how to create effective distractors for MC items. This study provides an extensive review of the published guidelines and findings from empirical research on developing distractors.

The Relationship between Experiences with Technology and Reading Performance in NAEP

Bitnara Park, American Institutes for Research; Young Yee Kim, American Institutes for Research; Jiao Yu, American Institutes for Research

Relationships between student experiences/exposure to technology and performance on recent NAEP assessments at grades 4 and 8 were examined. Results show that students with more exposures to technology tend to score lower on paper-based assessments. Additionally, the relationship between the nature of computer use and performance will be examined.

Friday, April 28, 2017
2:15 PM–3:45 PM, Salon M, Meeting Room Level, Paper Session, D9

Generalized Linear Mixed Models

A mixture IRT with a mixture response time model (mixIRT-RT)

Hye-Jeong Choi, University of Georgia; Allan Cohen, University of Georgia; Carrie Clark, University of Nebraska–Lincoln; Kimberly Espy, The University of Arizona

This study describes a mixture item response model with a mixture response time model by extending van der Linden's work. An important benefit of this model is that one can investigate relationship between accuracy and speed. A simulation study and an empirical example will be presented.

Matching bias of propensity score: can PSM be used for randomization purpose?

H Gary Cook, University of Wisconsin; Kyoungwon Bishop, University of Wisconsin

Despite its popularity, researches using propensity score matching often do not examine whether a matched sample conveys a population's distribution and how the results of treatment effects might introduce bias. This study illustrates how bias and imbalance exist in propensity matching using a large-scale English language proficiency assessment's data.

The Mixture Model with Internal Restrictions on Item Difficulty (MixMIRID)

Evan Olson, University of Maryland; Hong Jiao, University of Maryland

In this study, a proposed model, the mixture MIRID (MixMIRID) will be investigated. While mixture modeling has been applied extensively to item response theory (IRT) Rasch models (von Davier & Carstensen, 2007), modeling student subpopulations using the item dependencies of the MIRID has not been established.

Using item response models to handle measurement error in regression discontinuity designs

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

Assignment variables in regression discontinuity (RD) designs are often test scores calculated from categorical response variables. We propose latent variable regression models to obtain unbiased estimates of the treatment effects in RD designs. Both single- and multi-stage estimations are evaluated and contrasted to the observed-variable regression via Monte Carlo simulations.

Friday, April 28, 2017
2:15 PM–3:45 PM, Salon J, Meeting Room Level, Electronic Board Session: GSIC Graduate Student Poster Session, D10

Graduate Student Issues Committee

Brian Leventhal, Chair

Evelyn Johnson; Dubravka Svetina; Abeer Alamri; Maria Bertling; Brittany Flanery Crawford; David Martinez Alpizar; Rich Nieto

Electronic Board #1

An Investigation of Item Sensitivity to Response Style in Large-scale Assessments

SIEN DENG, University of Wisconsin-Madison; Daniel Bolt, University of Wisconsin-Madison

This study examines the varying sensitivity of rating-scale instruments to response style (RS) using a multidimensional nominal response model that allows varying RS discrimination parameters across items. Different rating scale types (agreement-type versus frequency-type scales) from PISA and PIRLS are found to have varying sensitivities to RS.

Electronic Board #2

Effects of Inclusion of a "Don't Know" Option on a Cognitive Test

Derek Sauder, James Madison University; Christine DeMars, James Madison University

The effects of including or not including a "Don't know" response option on a cognitive test were examined. The number of selections of "Don't know" versus number of items omitted, differences in average test scores, and gender or effort differences were of primary focus.

Electronic Board #3

Does Context Personalization of Mathematics Word Problems Reduce Rapid Guessing?

Audra Kosh, University of North Carolina at Chapel Hill

Using item response time data, this study sought to determine whether personalization of mathematics word problems (i.e., matching the problem context to a student's interest and using the student's name in the item stem) reduces effortless rapid guessing in a low-stakes assessment embedded in an online summer mathematics program.

Electronic Board #4

A Comparison of Categorical CFA and IRT for Examining DIF

Hwanggyu Lim, University of Massachusetts Amherst; Scott Monroe, University of Massachusetts Amherst

The purpose of this study is to compare DIF detection in IRT and CFA for categorical data, in particular with mixed-format tests by extending the work of Kim and Yoon (2011). The goal is to provide more specific guidelines for conducting DIF studies using either IRT or CFA.

Electronic Board #5

Extreme response style: which model is best?

Brian Leventhal, University of Pittsburgh; Clement Stone, University of Pittsburgh

More robust and rigorous psychometric models, such as IRT models, have been advocated for survey applications. However, item responses may be influenced by construct-irrelevant variance factors such as preferences for extreme response options. Through simulation methods, this study evaluates the use of IRT models designed to account for ERS.

Electronic Board #6

Wider Contexts for a Diverse Population: The Redesigned SAT's Reading Comprehension Test

Maryam Pezeshki, Georgia Institute of Technology; Susan Embretson, Georgia Institute of Technology

The reading sections of the old and redesigned SATs were compared using the Tools for the Automatic Analysis of Cohesion (TAACO) and of LExical Sophistication (TAALES). Results of MANOVA indicated that the redesigned test did not differ significantly from the old test in either word frequency or cohesion.

Electronic Board #7

Bootstrap Standard Errors of MIRT Equating

Stella Kim, The University of Iowa; Won-Chan Lee, The University of Iowa; Michael Kolen, The University of Iowa

This study is aimed to investigate bootstrap standard errors of various MIRT equating methods for mixed-format tests, under the random groups and common-item nonequivalent groups designs. For comparative purposes, several other procedures are also considered including the traditional equipercentile with presmoothing and two UIRT equating methods.

Electronic Board #8

Asymmetric Item Characteristic Curves and Misestimation of Item Information

Sora Lee, University of Wisconsin-Madison; Daniel Bolt, University of Wisconsin - Madison

The presence of asymmetric item characteristic curves (ICCs) can substantially affect ICC slopes and thus the information of an item. We examine in this paper how item information and the location of maximum information are misestimated when real asymmetry is ignored using real datasets and simulation analyses.

Electronic Board #9

Detect infeasibility in Automated Test Assembly

JIA MA, University of North Carolina at Greensboro

This project investigates how different level of mulcollinearity between attributes affects test assembly and to propose the method of detecting infeasibility problems by using multicollinearity diagnostics.

Electronic Board #10

A Guideline for Estimating Reliability Coefficients in Multilevel Confirmatory Factor Analysis

Abeer Alamri, Univeristy of South Florida; Eun SooK Kim, University of South Florida

This paper deliberates on the importance of examining level-specific reliability in multilevel confirmatory factor analysis (MCFA), and works as guideline on assessing and reporting reliability. It proposes the applicability and utility of five of most common MCFA reliability estimation procedures by presenting two demonstrations with simulation study and applied example.

Electronic Board #11

Using Principal Stratification to Assess Intervention Effectiveness at the Item Level

Nathaniel Raley, University of Texas at Austin; Adam Sales, University of Texas at Austin; John Pane, RAND Corporation

The Cognitive Tutor Algebra I (CTAI) curriculum has improved student outcomes in a randomized effectiveness trial. This study attempts a more nuanced analysis, introducing a new "potential outcomes" model for estimating the item-level causal effect of mastering relevant material using the Principal Stratification (PS) framework to elucidate this overall effect.

Electronic Board #12

Using Small Area Estimation to Rank Subpopulations on International Large-Scale Assessments

Luciana Cancado, University of Wisconsin-Milwaukee; Bo Zhang, University of Wisconsin-Milwaukee

This study investigates the accuracy and precision of Small Area Estimation (SAE) methods for state-level estimation and ranking based on country-level data using a sub-sample of the 2012 PISA survey from Brazil. Our preliminary results show that SAE estimates are consistent with full sample and official rankings but lack precision.

Electronic Board #13

Comparison of Bi-factor and Unidimensional IRT Model in Reading Comprehension Test

XINYUE LI, PENN STATE UNIVERSITY

Full information bifactor analysis is an important statistical method in psychological and educational measurement. Comparing the parameter estimates in bifactor model with those from the unidimensional model, results show that fitting a unidimensional IRT model to multidimensional data distorts the item parameters and bifactor model represents better model fit.

Electronic Board #14

Understanding PISA Problem Solving Assessment: Cross-country and Cross-race Comparisons

Shuang Wang, University of Wisconsin-Milwaukee; Hotaka Maeda, University of Wisconsin-Milwaukee; Bo Zhang, University of Wisconsin-Milwaukee

Problem solving assessment was first administered by PISA in 2012. We study its predictive validity by exploring its relationship with regular assessments and its measurement invariance as to the cognitive processes across countries and races. Finally, scores for each cognitive process are compared among the countries and races.

Electronic Board #15

Comparing a New Variable Compensation MIRT Model and Existing MIRT Models

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

The purpose of this study is to evaluate and interpret a variable compensation multidimensional item response theory (MIRT) model that allows for transformation between different correlation structures. In simulation, the new model is compared with existing MIRT models. It shows great flexibility and recovered parameters well with its rotation ability.

Electronic Board #16

Item nonresponse in large scale assessment data and software comparison

Taeyoung Kim, SUNY at Buffalo

Due to the low-stake nature of international large scale assessments, it is very likely that we encounter considerable amount of missing values. In this regard, it would be informative to see how accurately different software produce estimates in the presence of missing responses in those international surveys.

Electronic Board #17

Investigating Potential Source Caused Gender DIF in Reading Literacy Assessment

Yue Yin, University of South Florida; Yi-Hsin Chen, University of South Florida; Zhiyao Yi, University of South Florida

We investigated gender differential item function in reading comprehension tests and to further explore if reading purpose, comprehension process, and item type were the potential sources of gender DIF. Results from logistic regression indicated that 6 out of 24 items exhibited DIF but three sources of interest were not significant.

Electronic Board #18

Evaluation of Statistical Matching Methods: A Simulation Study Using NELS Data

Xin Yuan, Indiana University Bloomington

This study tries to explore how statistical methods work when basic assumptions hold and when they are violated through a simulation study. The results indicate estimates can be seriously biased when conditional independence assumption is violated. The validity levels of different matching methods were also tested after matching.

Electronic Board #19

Impact of Rater Severity on Essay Grade Population Invariance: A Simulation Study

Andrew Iverson, Washington State University; Brian French, Washington State University

This proposed study examines how essay raters varying in severity influence population invariance of different sized, but equally able, subgroups utilizing simulation. Conditions include the sampling method utilized for essay writer and raters, the disparity in and the size of subpopulations, and the number of raters in the model.

Electronic Board #20

Can Subscore Performance Predict Future Test Success?

Peter Ramler, University of Kansas; Matthew Schultz, American Institute of Certified Public Accountants

There is substantial discussion about the reasons for providing – or not – subscore information to test takers. Proponents of subscores note that scores provide usable diagnostic information regarding test takers proficiency. This study considers the usefulness of subscores in predicting retest success for repeat test takers.

Electronic Board #21

A Comparison of Multidimensional Item Response Item Parameter and Examinee Ability Estimation

Tianna Sims, Georgia State University

Advances in computational ability have produced new software which can estimate multidimensional item parameters and examinee abilities. This study will investigate the capability of various estimation techniques to recover multidimensional item parameter and examinee ability under differing conditions including test length, sample size, correlation, and test structure.

Electronic Board #22

Classifying Polytomous DIF in Small Sample Sizes using the Liu-Agresti Estimator

Elizabeth Patton, University of North Carolina Greensboro

This study evaluated the Liu-Agresti estimator of the cumulative common odds ratio when classifying the degree of polytomous differential item functioning (DIF) in small samples. Results indicate that in small samples the Liu-Agresti estimator underestimates true DIF for moderate to difficult items and overestimates true DIF for less difficult items.

Electronic Board #23

Effect of Priors on Item Parameter Estimation under the Graded Response Model

Shumin Jing, University of Iowa; Won-Chan Lee, University of Iowa

The purpose of this study is to evaluate the effect of various prior specifications on the accuracy of item parameter estimation under the graded response model. Considerable improvements are observed when the prior is properly specified.

Friday, April 28, 2017
4:05 PM–6:05 PM, Salon A, Meeting Room Level, Coordinated Session, E1

Can we estimate all US school test-score distributions from ordinal proficiency data?

Session Organizer: Andrew Ho, Harvard Graduate School of Education
Session Chair: Sean Reardon, Stanford University
Session Discussant: Henry Braun, Boston College

This symposium builds on previous work extending the "Heteroskedastic Ordered Probit" (HETOP) model to educational applications (Reardon, Shear, Castellano, & Ho, in press). That work demonstrated that the HETOP model can provide estimates of group means and standard deviations of latent achievement from coarsened proficiency data, such as those in categories like Basic, Proficient, and Advanced. The HETOP model enabled a national dataset of district-level data to be released to the public for interactive online visualization (Rich, Cox, & Block, 2016) and research (Reardon et al., 2016). However, our research identified shortcomings of the maximum likelihood HETOP estimator, specifically when sample sizes are small or the number of groups is very large. This posed a challenge for our aim to produce a school-level version of this national dataset. This symposium presents new methods of estimating the moments of continuous distributions that underlie ordinal data using the HETOP model, including a pooled data maximum likelihood estimator and a Bayesian estimator. We apply these approaches to estimate the moments of school test score distributions throughout the US. We implement and evaluate a cross-state linking method using NAEP. The result will be a publicly available national dataset of school-level average achievement.

Using pooled heteroskedastic ordered probit models to improve small-sample estimates
Benjamin Shear, University of Colorado Boulder; Sean Reardon, Stanford University

Flexible Bayesian models for inferences from coarsened, group-level achievement data
J. R. Lockwood, Educational Testing Service; Katherine Castellano, Educational Testing Service; Benjamin Shear, University of Colorado Boulder

Linking U.S. school test score distributions to a common scale
Andrew Ho, Harvard Graduate School of Education; Sean Reardon, Stanford University; Demetra Kalogrides, Stanford University

The distribution of academic achievement across US schools
Erin Fahle, Stanford University; Sean Reardon, Stanford University

Friday, April 28, 2017
4:05 PM–6:05 PM, Conference Room 1&2, Meeting Room Level, Coordinated Session, E2

Issues in Human Rater Calibration: How Much is Enough?

Session Chair: Cathy Wendler, Educational Testing Service
Session Discussant: Rosemary Reshetar, The College Board

The use of constructed response (CR) item types continues to grow and new task and item types utilizing human raters and/or automated scoring engines continue to be developed. While some of these tasks and item types lend themselves to automated scoring, human raters are likely to remain an integral part of CR scoring in the near future. A central challenge in scoring CR items is keeping the scoring criteria and ratings produced by human raters from shifting across or within scoring sessions. Rater calibration is often used as a way to help control rater shift. During the calibration process raters score a set of CR responses that have been predetermined to reflect various scoring categories. Raters who fail to meet minimum performance standards as part of calibration are often not allowed to perform operational scoring. The four papers in this session will discuss various efforts related to the rater calibration process for written CRs. In particular, they will address specific issues that may impact the efficiency of rater calibration and ultimately, the quality of CR scoring.

The Rater Calibration Process
Nancy Glazer, Educational Testing Service

Frequency of Calibration for Constructed Response Scoring
Bridgid Finn, Educational Testing Service; Kathryn Pedley, Educational Testing Service; Cathy Wendler, Educational Testing Service

Defining Optimal Calibration Sets
Isaac Bejar, Educational Testing Service; Brent Bridgeman, Educational Testing Service

Rater Scoring Accuracy Across a Multiple-Day Scoring Window
Cathy Wendler, Educational Testing Service; Fred Cline, Educational Testing Service

Friday, April 28, 2017
4:05 PM–6:05 PM, Conference Room 3&4, Meeting Room Level, Coordinated Session, E3

Looking Back and Moving Forward on Score Reporting Research and Practice

Session Chairs: Mary Roduta Roberts, Department of Occupational Therapy, Faculty of Rehabilitation Medicine, University of Alberta; Chad Gotch, Washington State University
Session Discussant: John Behrens, Pearson

As the primary interface between test developers and multiple educational stakeholders, score reports are a critical component to the success (or failure) of any assessment program. Relative to the vast resources devoted to test development, administration, and security, and the development of educational accountability policy in many large-scale testing contexts, score reporting has received much less attention. To maintain the integrity of assessment programs, attention to score reporting, including its antecedents and outcomes, is essential. The last 20 years have brought achievements in score reporting scholarship including the articulation of practice guidelines and methods for examining subscore added value. However, there are ongoing concerns regarding the extent to which information communicated within score reports is used and interpreted as intended. In the context of such advances in score reporting and heightened anxiety about testing programs, now is an ideal time to examine our progress and set a vision for meaningful advances. This innovative session adopts the perspectives of What has been, What is, and What could be? to position score reporting research as being more on par with other areas of measurement and education research.

A Systematic Review of Recent Empirical Research on Individual-Level Score Reports
Chad Gotch, Washington State University; Mary Roduta Roberts, University of Alberta

A Review and Evaluation of Changes in Score Reporting
Francis Rick, University of Massachusetts, Amherst; Yooyoung Park, University of Massachusetts, Amherst

Evaluating Score Report Effectiveness: Do Better Designed Reports Result in Better Outcomes?
Timothy O'Leary, Melbourne Graduate School of Education; John Hattie, Melbourne Graduate School of Education; Patrick Griffin, Melbourne Graduate School of Education

Communicating Results from a Competency-Based Performance Assessment
Mary Roduta Roberts, University of Alberta; Chad Gotch, Washington State University;Karin Werther, University of Alberta

Friday, April 28, 2017
4:05 PM–6:05 PM, Salon B, Meeting Room Level, Coordinated Session, E4

New Perspectives on Performing Job Analysis

Session Organizer: Adam Wyse, The American Registry of Radiologic Technologists
Session Chair: Matthew Burke, NCCPA
Session Discussant: Chad Buckendahl, ACS Ventures

A critical part of developing legally-defensible credentialing exams comes from using job analysis methods to define the content that should be included on the exam. This session provides new perspectives on several key components of conducting a rigorous and legally-defensible job analysis for credentialing programs. This includes looking at how the properties of a survey scale may impact job analysis results when using task inventory surveys, examining how the methods used to determine content weights may impact the weighting of content in the content specifications, discussing the role of competency modeling and how it can be modified to improve job analysis and resulting test specifications, and investigating the extent to which governing bodies overseeing credentialing programs are involved in the planning and selection of job analysis methodology. Commentary and discussion from a nationally known expert in credentialing provides links to recent court cases in which job analyses have been used to the support the use of credentialing exams for making decisions about examinees.

What is in a Survey Scale? Comparing Job Analysis Survey Scales
Adam Wyse, The American Registry of Radiologic Technologists; Ben Babcock, The American Registry of Radiologic Technologists; Dan Anderson, The American Registry of Radiologic Technologists; Carol Eckerly, Alpine Testing Solutions

Weightlifting for Strong Exams: Turning Rating Data into Exam Content Weights
Ben Babcock, The American Registry of Radiologic Technologists

Competency Modeling, Job Analysis, and Test Design for Credentialing Tests
Mark Raymond, National Board of Medical Examiners

The Policy Implications of Job Analysis Methodology
Jerry Reid, The American Registry of Radiologic Technologists

Friday, April 28, 2017
4:05 PM–6:05 PM, Salon C, Meeting Room Level, Coordinated Session, E5

Walking a tightrope: Navigating the balance of policy and psychometrics

Session Organizer: Susan Davis-Becker, ACS Ventures
Session Discussant: Gregory Cizek, University of North Carolina at Chapel Hill

Education and credentialing agencies regularly seek the assistance of outside expertise in the development, administration, evaluation, and monitoring of their program. In such roles, consultants often find themselves in challenging positions when they are asked to help a program meet particular policy goals or guidelines while adhering to psychometric best practices for test development which are in conflict with one another. The purpose of this session is to explore various roles that a psychometrician can engage in with a testing agency and review the types of challenges faced when trying to meet competing demands of policy and psychometrics. Each of four presenters will discuss a different type of psychometric role and examples of professional challenges that may be encountered in fulfilling that role. The presenters will share case studies (examples) of a situation that he or she faced in this role, how the psychometric and policy elements were misaligned, and how he or she navigated the situation to help the testing agency meet their goals and also implement a solution that was psychometrically defensible. These presentations, case studies, and judgmental processes will be reviewed by a discussant who has also fulfilled such roles across a number of programs.

Included but Independent: The role of a psychometric consultant
Lisa Keller, University of Massachusetts Amherst

Management of everything: The role of a testing company leader
Marianne Perie, Center for Educational Testing and Evaluation

The super advisors: The role of a TAC member
Chris Domaleski, National Center for the Improvement of Educational Assessment

Outside looking in: The role of an external evaluator
Susan Davis-Becker, ACS Ventures

Friday, April 28, 2017
4:05 PM–6:05 PM, Salon D, Meeting Room Level, Paper Session, E6

DIF session 1

A Local DIF Method to Evaluate Scoring Shift for Mixed-format Tests

Xuan (Adele) Tan, Educational Testing Service

A local DIF method to evaluate CR scoring shift for mixed-format tests is investigated. Using MC common items as the matching variable, STD P-DIF is calculated on total scores to evaluate performance differences indicating scoring shift. Outcome would be a reasonable evaluation criterion to detect scoring shift that warrants adjustment.

Accuracy of MH Methods and Wald Test to Detect DIF in CDMs

Hueying Tzou, National University of Tainan; Pei-Ming Chiang, National University of Tainan

The study was to investigate the efficacy of different DIF detection procedures in the cognitive diagnostic models (CDMs). We manipulated different attribute mastery profile distributions, CDMs, the ratios of sample size, and the number of attributes to examine the Type I error rate and power of three DIF detection procedures.

Comparison of Lasso Constraint Multiple Group Approaches for Detecting Differential Item Functioning

Jonathan Rollins III, The University of North Carolina at Greensboro; John Willse, The University of North Carolina at Greensboro

Two recently developed approaches (extensions of logistic regression and the Rasch model) for detecting differential item functioning (DIF) using lasso constraints are compared in a simulation study. Six conditions are manipulated across four groups: sample size, percentage of DIF items, DIF amount, impact via ability, DIF type, and ability distribution.

Demonstration of Multiple-Factor Multiple-Group NCDIF

Theresa Dell-Ross, Georgia State University; T. Chris Oshima, Georgia State University; Keith Wright, The Enrollment Management Association

In Multiple-Group NCDIF (MG-NCDIF), subgroups are compared against a "base" sample from the entire group of examinees. MG-NCDIF can be extended to include multiple factors (e.g., gender and race) simultaneously in the DIF analysis. This paper will demonstrate how Multiple-Factor Multiple-Group NCDIF (MFMG-NCDIF) can shed additional light on DIF analysis.

Deviation Coding within The Framework of Generalized Logistic Regression DIF Method

YUXI QIU, UNIVERSITY OF FLORIDA

For large-scale assessments that aim to compare educational achievement among subgroups within a country or across nations, consideration of test fairness is critical. This study presented and demonstrated the feasibility of a new coding scheme—deviation coding within the framework of generalized logistic regression DIF method.

Investigating the Effect of Differential Item Functioning on Proficiency Classification

Logan Rome, University of Wisconsin-Milwaukee; Bo Zhang, University of Wisconsin-Milwaukee

This simulation study aims to examine how the presence of DIF items may impact proficiency classification at various decision points. Of special interest are testing conditions in which DIF may go undetected by current methods but has a practical impact on the classification accuracy.

Friday, April 28, 2017
4:05 PM–6:05 PM, Salon K, Meeting Room Level, Paper Session, E7

Test Design issues with Linking

Session Chair: Ting Wang, ETS

An Investigation of Linking Invariance in Subgroups from Different Testing Modes

Hyeonjoo Oh, ETS; Hanwook Yoo, ETS; Shameem Gaj, ETS; Junhui Liu, ETS

This study investigates linking invariance in paper-based and online-based testing modes across multiple subpopulations (e.g., gender, race/ethnicity, and SES). The preliminary results suggest that there is very slight linking variance found by testing mode. Further investigation of linking invariance of subpopulations will be presented in the final paper.

Comparing Four Linking Methods for a High-Stakes Clinical Performance Examination

Alix Clarke, University of Alberta; Andrea Gotzmann, Medical Council of Canada; Fang Tian, Medical Council of Canada; André De Champlain, Medical Council of Canada; Sirius Qin, Medical Council of Canada

Linking of performance assessments is essential to account for differences in test form difficulty and ensuring accuracy of high-stakes decisions. Four different linking methods are compared in the context of a Canadian medical licensing clinical examination. Results inform the most appropriate application of linking methods to performance assessment testing programs.

Effects of Large Item Parameter Drift on Linking in Computerized Adaptive Testing

Xiaoran Li, University of Connecticut; Siang Chee Chuah, College Board; Melinda Montgomery, College Board

Anchor items are used to link pilot items onto the established scale. However, it can become problematic if there are items with large Item Parameter Drift (IPD). This study investigated the impact of large IPD on pilot item parameter estimation under different linking designs for a Computerized Adaptive Test (CAT).

Examining the Impact of Speeded Responses on Equating Results

Hongwook Suh, ACT; Sonya Powers, ACT; JP Kim, ACT

When examinees have insufficient time to complete all test items, speededness functions as an additional dimension. Previous research has found speededness to reduce the stability of item statistics and to impact equating results. This study uses empirical data to investigate how various degrees of speededness impact equating results.

Exploring item contexts and automated feedback as sources of item difficulty shifting

Ting Wang, ETS; Vinetha Belur, ETS; Ou Lydia Liu, ETS

Item context is one of many factors that influences performance on constructed-response (CR) tests. This study was conducted to examine the extent to which item characteristics influence student performance by the effects of immediate automated feedback on shifts in item difficulty. Item difficulty decreased before and after feedback was provided.

Robust IRT Scaling: Considerations in constructing item bank from tests across years

Jungnam Kim, NBCE; Dong-In Kim, DRC; Furong Gao, Pacific Metrics

This study investigates the impact of three different IRT scaling and equating methods in building an item bank of tests from 23 years of a national licensure exam. The study focuses on several key psychometric issues including scale drift and equating errors.

Friday, April 28, 2017
4:05 PM–6:05 PM, Salon L, Meeting Room Level, Paper Session, E8

Test Design isssues with Diagnostic Classification Models

Session Chair: Hua-hua Chang, University of Illinois at Urbana-Champaign

Cognitive Diagnosis in Classroom: The General Nonparametric Classification Method

Chia-Yi Chiu, Rutgers, The State University of New Jersey; Yan Sun, Rutgers, The State University of New Jersey; Yanhong Bian, Rutgers, The State University of New Jersey

Cognitive Diagnosis (CD) performs well for large-scale assessment systems; but in small-scale settings, the performance of CD is usually poor because the commonly used parametric methods fail when sample sizes are small. A general nonparametric classification method (GNPC) is proposed that allows CD to be used in small educational settings.

Conditional Classification Accuracy and Consistency of Cognitive Diagnosis Models

Charles Iaconangelo, Rutgers, The State University of New Jersey

As the accuracy and consistency of an assessment are an essential part of the validity argument, indices are proposed that estimate the accuracy and consistency conditional on the latent class. Compared to the alternative, parametric Monte Carlo approaches, the indices provide easy-to-compute estimates that are close to the empirical values.

Item Selection Method for Attribute Hierarchical DCM-CAT

Yu Bao, University of Georgia; Laine Bradshaw, University of Georgia

Assessments of attribute hierarchies typically result in disproportionate statistical information for attributes at different hierarchy locations. Leading DCM-CAT item selection methods administer sub-optimal items when item pools contain unequal information for attributes. We propose an attribute-level item selection algorithm for adaptive DCMs that is appropriate when hierarchies are present.

Misspecification of Attribute Structure in Diagnostic Measurement

Ren Liu, University of Florida

This study provides a framework for understanding misspecifications of attribute relationships. It investigated the effects of attribute structure misspecification on model fit, item fit, and respondent classification accuracy. Results show that various types of misspecifications from external shapes or internal organizations have different impacts on the outcomes of interest.

Multiple-Choice Cognitive Diagnosis Model with Computerized Adaptive Testing

Hulya Duygu Yigit, Rutgers, The State University of New Jersey; Miguel Sorrel, Universidad Autónoma de Madrid; Jimmy de la Torre, The University of Hong Kong

Several item selection methods have been proposed for CD-CAT working with dichotomous data. Jensen-Shannon divergence (JSD) index can be used to fill the gap for CD-CAT applications with polytomous data. Based on the simulation study, the performance of JSD is evaluated under the multiple-choice deterministic inputs, noisy "and" gate model.

Using Cognitive Diagnostic Computerized Adaptive Testing to Help Classroom Learning

Hua-Hua Chang, University of Illinois at Urbana-Champaign; Hyeon-Ah Kang, Columbia University; Susu Zhang, University of Illinois at Urbana-Champaign

The objective of the current research is to demonstrate that CD-CAT can improve classroom learning by an ongoing case study at the University of Illinois at Urbana-Champaign, where a CD-CAT system has been developed to help low-performing students in a difficult undergraduate physics course to increase the course retention rate.

Friday, April 28, 2017
4:05 PM–6:05 PM, Salon M, Meeting Room Level, Paper Session, E9

Applications of Multilevel Modeling

An Empirically-Derived Index of High School Academic Rigor
Jeff Allen, ACT, Inc.; Edwin Ndum, ACT, Inc.

Methods for measuring high school academic rigor are advanced by: 1) optimizing the prediction of first-year college GPA using nominal coding of high school coursework and grades, and 2) incorporating 8th grade test scores and grades to isolate the effects of high school coursework and grades.

Applications of Latent Class Growth Models for Educational Research and Practice
Anthony Fina, Iowa Testing Programs, University of Iowa

This paper demonstrates the utility of latent class growth models through several practical applications. The emphasis is not model building, but using the defined models to assess student learning and to inform decisions in both K-12 and higher education. For each application, impact of status versus growth measures are compared.

Cross-Sectional and Longitudinal Profiles Estimated from the Early Child Longitudinal Study
Andrea McNamara, Fordham University; Se-Kang Kim, Fordham University

This paper conducted profile analysis for a large-scale longitudinal educational study, utilizing principal component analysis. Cross-sectional and longitudinal profiles were extracted and relationships between them were explored. Such analysis is advantageous for both short-term and long-term decision-making by teachers/parents and policy makers.

Hierarchical Structures in CD-CAT Applications
Mehmet Kaplan, The Turkish Ministry of National Education; Lokman Akbay, Rutgers, The State University of New Jersey

This study examines the impact of hierarchical structures on the performance of item selection indices, and also demonstrates the improvement in classification accuracy when the hierarchical structures in attribute vectors are considered. The results showed that considering the prior distribution of the attribute vectors improved the classification accuracy.

Identifying Profiles of Reading Performance Based on Component Measures
Jonathan Weeks, Educational Testing Service; John Sabatini, Educational Testing Service

This study examines student profiles of reading based on measure of six separate component skills, using data for students in grades 6 – 9. Classifications based on discretized versions of the component scores and hierarchical cluster analysis are used to identify instructionally relevant profiles of reading performance.

The Effects of Test-based English Language Learner Classification Criteria on Academic Achievement
Nami Shin, CRESST/UCLA

This study examines how different sets of initial English Language Learner classification criteria have impact on subsequent students' academic achievement. Using longitudinal data from a large school district, this study compares two groups of students who were initially classified by two different sets of classification criteria.

Friday, April 28, 2017
4:05 PM–6:05 PM, Salon J, Meeting Room Level, Electronic Board Session, Paper Session, E10

Electronic Board #1

The Impact of Increased Item Exposure on Difficulty and Pass Rates

John Sessoms, University of North Carolina at Greensboro; Jerome Clauser, American Board of Internal Medicine

Item exposure may result in easier items. If all items become easier, this difficulty change cannot be detected and will result in higher, yet inaccurate test scores. Compared to low-exposure items, high-exposure items became much easier from the first to second administration. Test scores and pass rates were meaningfully affected.

Electronic Board #2

Performance of Raschtree for Uniform DIF Detection with Continuous Covariates

Holmes Finch, Ball State University; Julianne Edwards, Azusa Pacific University; Brian French, Washington State University; Carolin Strobl, University of Zurich

Assessment of differential item functioning (DIF) is necessary to provide score validity evidence. A recently developed method, Raschtree, may prove particularly useful for assessing DIF with multiple covariates, some of which are continuous. This study examined the performance of Raschtree with continuous DIF covariates and found that it worked well.

Electronic Board #3

Methods for handling zero expected cell frequency in contingency table based statistics

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

When calculating chi-square statistics, pooling expected cells to achieve a minimum frequency is common. Several values of minimum expected frequencies in calculating Orlando and Thissen's (2000) item level S-X2 and S-G2 fit statistics are examined in this simulation study. Distributional properties, type I error, and power are evaluated.

Electronic Board #4

Interruption estimation of online testing: Maximum Likelihood approach

Kyoungwon Bishop, University of Wisconsin; H Gary Cook, University of Wisconsin

One challenge of the advent of large-scale online assessment is how to understand and evaluate the impact of interruptions on student test performance. This presentation shares a study of how the WIDA Consortium used a maximum likelihood statistical technique to identify the impact of interruptions on English learners' test scores.

Electronic Board #5

Rethinking Content Validation: A Call for Increasing Rigor on Alignment Evidence

Catherine Welch, University of Iowa; Stephen Dunbar, University of Iowa; Anthony Fina, University of Iowa; Yi Gui, University of Iowa

Alignment results are used as evidence of content validity and test appropriateness, although few studies address the adequacy of results for the purposes they are intended to serve. This paper uses generalizability theory to develop definitions of adequacy and document the degree to which alignment results can produce interpretable results.

Electronic Board #8

Detection of Test Speededness Using Change-Point Analysis with Response Time Data

Can Shao, National Board of Osteopathic Medical Examiners; Ying Cheng, University of Notre Dame

A test is speeded when examinees don't have time to fully consider every item on the test. We propose to detect speededness using change-point analysis with response time data. Simulation shows that the procedure is efficient in detecting both speeded examinees and the point at which they start to speed.

Electronic Board #9

A Statistical Criterion to Assess Fitness of Cubic-spline Postsmoothing

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

Log-linear presmoothing methods uses various statistical criteria to select the optimum degrees of smoothing. However, for cubic-spline postsmoothing, there is no current statistical criterion; visual inspection has been an important tool in choosing such degree. This study introduces a new statistical criterion for assessing the fitness of cubic-spline postsmoothing method.

Electronic Board #11

Deriving Rapid Response Thresholds for Investigating Test Speededness

Richard Feinberg, National Board of Medical Examiners; Daniel Jurich, National Board of Medical Examiners

Objective methods for determining rapid responding thresholds remain relatively unexplored. This study investigated a novel method employing the relationship between probability of correct response and response time to produce a simple equation for determining rapid responding thresholds at the item level. Applications and comparisons using large-scale operational data are presented.

Electronic Board #12

Investigating the relationship between ability and response time for computer-based tests

Shu-chuan Kao, Pearson

This study investigated the relationship between response time and Õ-b by correct and incorrect responses. Initial results concluded that the relationship between Õ-b and response time is negative for the correct response in the unspeeded section but mostly positive regardless of correct or incorrect responses in the speeded section.

Electronic Board #13

Fixed Item Parameter Calibration with MMLE-EM Using a Fixed Ability Prior

Sung-Hyuck Lee, ACT; Hongwook Suh, ACT

A new fixed item parameter calibration method (FIPC) with a fixed ability prior is proposed. This approach outperforms existing FIPC methods in estimating item parameters on underlying ability scale in that they are less biased and more stably estimated in the presence of new items with poor model fit.

Electronic Board #14

A Preliminary Study on Mixed Membership with Rash Model

Guoguo Zheng, University of Georgia; Hye-Jeong Choi, University of Georgia; Brian Bottge, University of Kentucky

We applied a mixed membership model with Rash model for students' response in a fractions computation test. Unlike a mixture IRT, this model allows students to have partial latent class membership and to switch classes across test items. The model is estimated using MCMC in R.

Electronic Board #15

Are computer-based writing tests comparable to paper-based tests for primary students?

Lucy Lu, Department of Education, New South Wales, Australia; Margaret Turnbull, Department of Education, New South Wales, Australia

This research investigates mode effects for writing assessments in the Australian context, by comparing writing samples produced by 3000 students on computer-based and paper-based tests. The study will examine if the mode effect differs across demographic groups and if keyboarding proficiency is a contributing factor to the mode effect.

Electronic Board #16

Scripted On-the-Fly Multistage Testing

Edison Choe, ACT; Bruce Williams, ACT; Sung-Hyuck Lee, ACT

Multi-stage tests (MST) have been proposed as an alternative to CAT, in which item review is allowed in each stage. In the present study, scripted on-the-fly MST was administered, in which content constraints and item exposure rates were controlled without diminishing reliability too much, as compared to a CAT.

Electronic Board #17

The Effect of Composite Scoring Methods on Classification Decisions

Nathan Minchen, Rutgers, The State University of New Jersey; Hao Song, National Board of Osteopathic Medical Examiners; Isaac Li, National Board of Osteopathic Medical Examiners; Qiongqiong Liu, National Board of Osteopathic Medical Examiners

Clinical decision making (CDM) items are designed to measure the clinical judgement of physicians using clinical scenarios with open-ended and multiple-choice multiple-correct questions. This study investigates approaches to constructing a composite score that integrates CDM items with conventional multiple-choice items on a medical licensing examination, and evaluates its psychometric contribution.

Electronic Board #18

A Group-level Wald Test for Response-based Cheating Detection

Zhuangzhuang Han, Teachers College, Columbia University; Xiang Liu, Teachers College, Columbia University; Matthew Johnson, Teachers College, Columbia University

Most of IRT-based methods for cheating detection focus on the individual level. Answer changes (ACs) and person-fit analysis statistically model the aberrance on individual response patterns. We propose a Wald test based on group-level latent traits. Derivation is shown and simulation study demonstrates its properties and effectiveness.

Electronic Board #19

Analyzing PROMIS Social Relationship Measures with Polytomous IRT models

Joseph Olsen, Brigham Young University; James Olsen, Renaissance Learning, Inc,

We use cumulative probability, adjacent category, and sequential IRT models, including models with category boundary discrimination parameters, to estimate calibration parameters for the PROMIS measures of social isolation, companionship, emotional support, instrumental support, and informational support. The paper applies a basic framework for organizing, estimating, and comparing the models.

Electronic Board #20

An application of multivariate generalizability theory to examine composite score reliability

Kelly Foelber, James Madison University; Amanda Clauser, National Board of Medical Examiners

The purpose of this study is to apply multivariate generalizability theory to examine the reliability of composite scores from a three-component high-stakes performance assessment under different component weighting schemes. The rationale for using multivariate generalizability theory and the appropriate interpretations of the resulting estimates will be discussed.

Electronic Board #21

Validating a Writing Self-Efficacy Scale Using a Large-Scale Testing Program

Tanesia Beverly, University of Connecticut; Jan Alegre, Educational Testing Service; Jonas Bertling, Educational Testing Service

A writing self-efficacy (WSE) scale was established using a large-scale testing program's pilot data. Hierarchical regression was employed to assess criterion-related validity. Initial results show a moderate relationship between scores on two writing essays and WSE. The addition of demographic variables explains roughly 33% of the variance in writing scores.

Electronic Board #22

A Validity Framework in Support of the Principled Use of Repurposed Assessments

Maria Oliveri, Educational Testing Service; Rene Lawless, Educational Testing Service

A validity framework describing the challenges, complexities, and suggestions for developing and using repurposed assessments (i.e., assessments developed for one population and later used with other linguistically or culturally diverse populations) will be presented. Examples from higher education assessments will be used to exemplify critical features of the framework.

Electronic Board #23

Measurement of Achievement Growth using Admissions Data: Implications for Admissions Processes

HyunJoo Jung, University of Massachusetts Amherst; KyeongJin Kang, Sogang University; Bomi Kim, Sogang University

We investigate whether different types of admissions processes – one based on the Korean College Scholastic Ability Test and another based on the admissions officer system – have an impact on the growth of student achievement using longitudinal admissions data from a university in Korea.

Friday, April 28, 2017
4:30 PM–6:30 PM

Graduate Student Social (Graduate Students Only)

Yard House (River Walk)
849 E. Commerce St.
San Antonio, TX 78205l

Friday, April 28, 2017
6:30 PM–8:00 PM, Salon I, Meeting Room Level

NCME and Division D Reception

Saturday, April 29, 2017
8:00 AM–10:00 AM, Salon EF, Meeting Room Level

NCME Breakfast, Presidential Address and Business Meeting

NCME Presidential Address
Mark Wilson
UC Berkeley, Berkeley, CA.

Join your friends and colleagues at the NCME Breakfast and Business Meeting at the Marriott Rivercenter. Theater style seating will be available for those who did not purchase a breakfast ticket but wish to attend the Business Meeting.

Saturday, April 29, 2017
10:35 AM–12:05 PM, Salon A, Meeting Room Level, Invited Session, F1

2017 NCME Award for Career Contributions to Educational Measurement
Dr. Linda L. Cook

Saturday, April 29, 2017
10:35 AM–12:05 PM, Conference Room 1&2, Meeting Room Level, Coordinated Session, F2

Student Learning Objectives and the Challenge of Campbell's Law

Session Chair: Derek Briggs, University of Colorado
Session Discussant: Andrew Ho, Harvard Graduate School of Education

Student Learning Objectives (SLOs) involve a process in which teachers establish measurable achievement goals for their students, assess students at the outset of an instructional period and then establish targets for student growth over the duration that period. Over the past five years they have been widely implemented as a means of fulfilling both formative and summative purposes. It is an open question whether SLOs are capable to surviving the challenge posed of Campbell's Law, which holds that the more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption or distortion. In this session we present results from two different multi-year evaluations of the SLO process. One evaluation comes from the implementation of SLOs in schools throughout a state; the other come from an implementation in the largest school district in Colorado. Two presentations focus attention on the variable quality of the SLOs teachers are enacting and aspects of the process that have made it difficult to integrate into their daily classroom practices. The other two presentations use empirical data to evaluate the validity of SLOs as a measure of teacher effectiveness.

Autonomy at a Price: Considering the Pit-falls of a Teacher-Driven Implementation Model
Elena Diaz-Bilello, University of Colorado; Amy Burkhardt, University of Colorado

Teacher and Leader Perceptions of Student Learning Objectives: A Case Study
Susan Lyons, National Center for the Improvement of Educational Assessment

Examining the Use of SLOs to Evaluate Teachers
Derek Briggs, University of Colorado; Rajendra Chattergoon, University of Colorado

The Validity of Teacher Scores Based on Student Learning Objectives
Katie Buckley, Transforming Education

Saturday, April 29, 2017
10:35 AM–12:05 PM, Conference Room 3&4, Meeting Room Level, Coordinated Session, F3

Mixture and Bayesian IRT Approaches to Modeling Guessing and Beyond

Session Chair: Lihshing Wang, University of Cincinnati
Session Discussant: Jean-Paul Fox, University of Twente

Guessing (or pseudo-guessing) in ability tests has been known to compromise both reliability and validity by introducing random and systematic errors into observed test scores. While the 3-parameter item response theory models provide appealing alternatives to the traditional non-IRT methods for the correction for guessing, the c parameter representing the lower asymptote of the item response curve has been widely criticized for its conceptual ambiguity and computational inefficiency. This coordinated session revisits this long-standing issue by introducing recent advances in using the Bayesian and mixture approaches to modeling guessing behavior under both parametric and nonparametric frameworks. This session opens with a brief review of issues associated with the c parameter in the 3P IRT models, followed by a discussion of the pros and cons of mixture modeling to identify guessers, partial guessers, and non-guessers. The Bayesian approach is then introduced in various contexts, including using the latent ability or response time as prior and combining the Bayesian and mixture approaches to model partial guessing. The session concludes with a non-parametric extension to both item and person estimation beyond guessing. The panel discussion format promotes interactive dialogue and intensive exchange among the presenters, the discussant, and the participants.

Lihshing Wang, University of Cincinnati

Gregory Camilli, Rutgers University

Yanyan Sheng, Southern Illinois University

Jiaqi Zhang, University of Cincinnati

Bernard Veldkamp, University of Twente

Jing Cao, Southern Methodist University

Lynne Stokes, Southern Methodist University

Yisheng Li, University of Texas MD Anderson Cancer Center

Saturday, April 29, 2017
10:35 AM–12:05 PM, Salon B, Meeting Room Level, Coordinated Session, F4

Perspectives on Quality in Assessment Practice

Session Chair: Thanos Patelis, Center for Assessment
Session Discussant: David Williamson, Educational Testing Service

Educational measurement relies on standards and guidelines in the development and operations of quality assessments. Across the professional assessment community, the Standards for Educational and Psychological Testing serve as these guidelines. However, the actual quality of assessments and assessment products is driven by other aspects, as well, including the context, the intended and actual use, the stakeholders involved, the interpretation of the Standards, other guidelines, business interests and practices, etc. The purpose of this session is to offer various perspectives on the drivers of quality assessment practice. The first presenter will provide information about other standards and guidelines from outside the U.S. that provide valuable guidance. The second presenter will focus on how assessment programs can consider their testing purposes, and how they can be matched to the type of evidence required, particularly as program goals evolve. The third presenter will discuss issues associated with tests that have multiple uses and conflicting guidance about the appropriateness of those uses. The fourth presenter will provide examples of efforts that have been undertaken to evaluate the quality of assessments and assessment programs. Finally, the fifth presenter will incorporate standards and guidelines for quality assessments into the broader context of new product development.

International Standards for Educational and Psychological Testing
Stephen Sireci, University of Massachusetts-Amherst

The Use of the Test Standards with Evolving Program Uses
Andrew Wiley, ACS Ventures

Policy Makers, Peer Review, Standards, and Best Practices: Assessments Servicing Multiple Masters
Kevin Sweeney, The College Board

Efforts to Evaluate the Quality of Assessment Programs
Thanos Patelis, Center for Assessment

Saturday, April 29, 2017
10:35 AM–12:05 PM, Salon C, Meeting Room Level, Coordinated Session, F5

Stories Told by Test Cheaters: What we can learn from them

Session Chair: John Fremer, Caveon Test Security

Stories told by test cheaters are presented and analyzed. Each story was originally told to a test administrator in an operational testing session. The collection on which this session is based is from a publication of the National College Testing Association. The lead off presenter, Jarret Dyer, is the collector of the stories. After the stories are presented, the contents will be commented on from three perspectives:

- National Testing Programs
- Research and Test Program Management
- Test Security Company

Time will be managed so that attendees can also provide their perspective on the stories. The goal of each presentation will be to learn from each story and from sets of stories with common characteristics. How can we improve our research, development, and implementation policies, procedures, and materials by paying close attention to how cheaters view the testing situation as we can judge from their comments? What more would we like to know and how can we design and execute studies to find out?

Stories Told by Test Cheaters: What can we learn from them

Jarret Dyer, College of DuPage

Saturday, April 29, 2017
10:35 AM–12:05 PM, Salon D, Meeting Room Level, Paper Session, F6

Test design issues in classroom assessment

Junhui Liu, Educational Test Service

Junhui Liu, Educational Test Service; Shu-Ying Chen, National Chung Cheng University

The current simulation study is to evaluate the usage of minimum discriminant information adjustment (MDIA) to construct matched samples for DIF analyses in small sample data. The performance of MDIA DIF will be compared to MH and standardization (SMD) methods for DIF detection in the context of large-scale state assessments.

Building an Innovative Formative Assessment to Gather Evidence about Students' Argumentation Skills

Yi Song, Educational Testing Service; Jesse Sparks, Educational Testing Service

A game-based assessment informed by learning progressions was developed to gather evidence about middle school students' argumentation skills for formative classroom use, wherein players complete five argumentation activities around students' access to junk food. We present the assessment design, results from a pilot study (N=104), and implications for game-based assessment.

Science Teachers' Use of Assessment Accommodations for English Language Learner (ELL) Students

Min Li, University of Washington at Seattle

Given the rapid increase of ELLs in public schools, teachers' classroom assessments need to address their learning needs. Unlike earlier research that has primarily focused on ELL accommodations in large-scale testing, this study examines a sample of 32 science teachers' uses of accommodation strategies in written assessment tasks.

The relationship between response time and accuracy in a large scale assessment

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

This study investigates the relationship between response time and accuracy using data from a large scale assessment. A double centering technique is utilized to remove person and item effects in response time. Results reveal different relationships across disciplines and item difficulty levels. Findings will inform new model development.

Saturday, April 29, 2017
10:35 AM–12:05 PM, Salon K, Meeting Room Level, Paper Session, F7

Multidimensional estimation of subscores

Session Chair: Carl Setzer, AICPA

Are Diagnostic Profiles of Reading Skill Mastery Valid and Reliable?

Yi-Hsin Chen, University of South Florida; Isaac Li, University of South Florida; Elizabeth Shaunessy-Dedrick, University of South Florida; Robert Dedrick, University of South Florida

It is challenging to dissect a reading ability in cognitive diagnostic assessments. This study attempted to refine cognitive components to a reading section of the Iowa Test of Basic Skills (ITBS) and further provide diagnostic profiles. The evidence of reliability and validity of diagnostic profiles were also provided.

Evaluation of Ability Estimates under 2PL Bi-factor Testlet Model

Chalie Patarapichayatham, Southern Methodist University; Akihito Kamata, Southern Methodist Univesity

This study analyzed reading test data to investigate characteristics of factor score estimates under various specifications of the bi-factor testlet model. A simulation study conditions are developed from the real data analysis to investigate whether testlet-factor factor scores can be used as estimates of testlet subscale scores.

How much can we gain from collateral information for subscore reporting?

Xiaolin Wang, Indiana University - Bloomington; Dubravka Svetina, Indiana University - Bloomington; Shenghai Dai, Indiana University - Bloomington; Ou Zhang, Pearson

The current study aims to compare the performance of three augmentation subscoring methods – MIRT, Yen's OPI, and Wainer's augmentation - with their non-augmentation counterparts when collateral information of different qualities is available.

The Impacts of Psychometric Item Properties on Subscore Reliability in Multidimensional IRT

Yoonjeong Kang, American Institutes for Research; Youngmi Cho, Pearson; Ming Li, Georgetown University

This study systematically investigates the impact of subtest length, correlation between subtests, and item properties on subscore reliability in multidimensional item response model framework. The results show that item properties and subtest length are the most influential factors on the accuracy of subscore ability estimates and subscore reliability.

Using a Machine Learning Classifier to Generate Candidate Feedback

Carl Setzer, AICPA

Research on subscores has increased in recent years. One area from we can borrow is that of machine learning. This paper examines the utility of a naïve Bayes classifier at the content area level within a large-scale licensure examination. Different approaches to assessing the model quality are reviewed and applied.

Saturday, April 29, 2017
10:35 AM–12:05 PM, Salon L, Meeting Room Level, Paper Session, F8

Issues in Standard Setting 2

An Evaluation of Conjunctive, Disjunctive and Compensatory Standard-Setting Strategies through Predictive Validity

Xin Liu, Ascend Learning

The main purpose of this study is to gather predictive validity evidence attesting to the important decision which strategy to use in standard setting: conjunctive, disjunctive, or compensatory.

Cutscore Distribution Theory (CDT) for the Bookmark Standard Setting Procedure

Joseph Fitzpatrick, University of Kansas; William Skorupski, University of Kansas

A psychometrics of standard setting that accounts for panelist consistency and accuracy is extended to the Bookmark procedure. The effects of these factors on resulting cutscores are demonstrated using simulated and real data. Results are compared with traditional approaches, which model panelist consistency but not panelist accuracy.

Equating Standard-Setting Forms Using the Circle-Arc Method

Xiaoyu Qian, Educational Testing Service; Weiling Deng, ETS

We used the Circle-Arc equating on two standard-setting forms with small sample sizes in the proposal. The preliminary results suggest that, the Circle-Arc method is probably fairer and more efficient compared to the current operational practice. More replications of real test data simulation is needed to support the initial results.

Interval Validation Method: An Investigation of Interval Length and Item Pool Size

William Insko, Houghton Mifflin Harcourt; Stephen Murphy, 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 study interval length and item pool size. Several recommendations for selecting an optimal interval length and a strategy for systematically reducing item pools are discussed.

The Impact of Training on Judge Consistency for Angoff Standard Setting Exercises

Melissa Margolis, National Board of Medical Examiners; Brian Clauser, National Board of Medical Examiners

Although procedures used to train Angoff standard-setting judges have received significant attention, the extent to which these procedures actually improve the internal consistency of the resulting judgments has received little attention. This paper evaluates that aspect of training for operational standard-setting exercises implemented for the United States Medical Licensing Examination.

Saturday, April 29, 2017
10:35 AM–12:05 PM, Salon M, Meeting Room Level, Paper Session, F9

Score Reporting

An Examination of Design Considerations in Reporting Reliable Aggregate-Level Subscores
Usama Ali, Educational Testing Service; Joseph Rios, Educational Testing Service

Limited research has investigated the design considerations necessary for reporting reliable aggregate-level subscores. This study investigated three variables that may impact the reliability of such scores: a) the number of within-form subdomain items, b) the number of total test forms, and c) the sample size within a group.

Disentangling the Impressions of Achievement Given by Various Reporting Metrics
Robert Ankenmann, The University of Iowa; Stephen Dunbar, The University of Iowa; Catherine Welch, The University of Iowa

This study uses state-wide assessment data to examine changes in reading and mathematics achievement of students in Grades 3 through 5 between 2001 and 2011. Results based on three reporting metrics/indexes (percent proficient, average scale score, and growth) are compared at various levels of aggregation.

Enhancing the Use of Standardized Testing Performance via Benchmark Reports
Johnny Denbleyker, Houghton Mifflin Harcourt; Linlin Wu, University of Illinois-Chicago

This study compares and contrasts two methods that employ the beta-binomial model to calculate a robust and relative estimate of effect-size to produce a Benchmark Report graphic for a CAT. The report is argued to be useful for diagnosing curricular/instructional strengths and weaknesses at an aggregate level

Promoting Accurate Score Report Interpretation and Use for Instructional Planning
Meagan Karvonen, University of Kansas; Russell Swinburne Romine, University of Kansas; Amy Clark, University of Kansas; Jennifer Brussow, University of Kansas; Neal Kingston, University of Kansas

This presentation describes results from two studies on the interpretation and use of alternate assessment (AA-AAAS) score reports. The first study focuses on usability of report contents for communication with parents and instructional planning. The second study examines the impact of interpretation resources on educators' understanding of report contents.

Visualizing Effect Sizes Across the Full Distribution
Daniel Anderson, University of Oregon; Joseph Stevens, University of Oregon; Joseph Nese, University of Oregon

Data visualization is an integral part of analysis and communication of results. This paper focuses on visualizing achievement gaps, with an emphasis on effect sizes. We discuss and introduce methods for displaying achievement gaps across the full distribution, rather than at a single location on the scale (e.g., the means).

Saturday, April 29, 2017
10:35 AM–12:05 PM, Salon J, Meeting Room Level, Electronic Board Session, Paper Session, F10

Electronic Board #1

Bias vs. precision: the effects of retaining DIF items on scale scores

Jue Liao, University of California, Los Angeles; Megan Kuhfeld, The University of Texas at Austin; Mark Hansen, University of California, Los Angeles & CRESST

The impact of DIF on test scores is investigated through a simulation study with conditions chosen based on a literature review. Findings indicate conditions considered problematic in the DIF detection literature do not always lead to substantial test score bias, while dropping items with DIF can reduce overall score precision.

Electronic Board #3

Weighted Hamming Distance: A Simple Nonparametric Person-Fit Index

Xin Luo, Northwest Evaluation Association (NWEA); Jiahui Zhang, Michigan State University

This study developed a simple and intuitive index—weighted Hamming distance (WHD)—to detect aberrant response patterns. The aberrant response type, test length, and aberrant response proportion were manipulated to simulate various test settings. The result showed that WHD can maintain high detection power with less computation than the existing methods.

Electronic Board #4

The Comparative Performance of Tree-Based Machine Learning Algorithms in Statistical Software

Scott Wood, Pacific Metrics Corporation

Gradient-boosted machines (GBMs) are popular data modelling techniques in automated scoring engines and assorted educational research applications. A comparative study between two popular GBM implementations—R's gbm library and SciKit-Learn's GradientBoostedRegression object—shows that the two algorithms can yield different results when configured with the same parameter settings.

Electronic Board #5

The Importance of Mastery: Evidence from the Cognitive Tutor Effectiveness Study

Adam Sales, University of Texas College of Education; John Pane, RAND Corporation

The Cognitive Tutor online tool structures student work via mastery learning: students proceed after mastering prerequisite skills. However, teachers are able to override the mastery constraint. We use randomized trial data, psychometric modeling, and Principal Stratification to estimate the relationship between mastery learning CTAI treatment effect.

Electronic Board #6

Automated Formative Assessment

Joshua Wilson, University of Delaware; Rod Roscoe, Arizona State University - Polytechnic; Yusra Ahmed, University of Houston

This study examines whether automated assessment tools might facilitate formative writing assessment. Samples of writing from middle-grade students were analyzed using PEG Writing and Coh-Metrix. Hypothesized models were evaluated using confirmatory factor analysis. Results indicate that automated formative writing assessment may be a viable alternative to traditional human-scored assessment methods.

Electronic Board #7

Investigate Impact of Item Parameter Drift on Pretest Item Calibration in CAT

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

This study intends to investigate possible impact of items parameter drift on pretest item calibration under CAT design. Simulation conditions include drift types, drift magnitudes, drift conditions, percentages of drifted items, pretest item selection procedures, pertest item administering methods and different sizes of calibration sample for pretest items.

Electronic Board #8

Utilizing Skill Decay in the Design and Development of K-Career Assessments

Hope Clark, ACT; Pamela Paek, ACT

The purpose of this paper is to discuss the implications of skill decay for K-12, PSE and workforce assessment design and credential development, with recommendations and design/development considerations to create a more seamless approach to connecting learning from early classroom study to career.

Electronic Board #9

Searching for an Optimal Sample Size for a Rescored Sample in Equating

Chi-wen Liao, Educational Testing Service; Wei Wang, Educational Testing Service; Yi Cao, Educational Testing Service

This study investigated what a sample size for a rescored sample should be in mixed-format test score equating. The preliminary results indicated the smaller the rescored sample size, the greater the SEE (ranged from 0.50s to 1.40s). The sample size was not much related with equating bias.

Electronic Board #10

A Higher-Order Cognitive Diagnosis Model for Polytomous Attributes and Polytomous Responses

Peida Zhan, Beijing Normal University; Wen-Chung Wang, The Education University of Hong Kong; Lijun Wang, Zhejiang Normal University

Existing cognitive diagnosis models (CDMs) assume dichotomous attributes and dichotomous responses. In this study we developed new CDMs for polytomous attributes and polytomous responses. A higher-order structure was imposed on the polytomous attributes. A brief simulation demonstrated a acceptable parameter recovery by using the Bayesian methods for parameter estimation.

Electronic Board #11

Dealing with Missingness in Cognitive Diagnostic Models When the Q-Matrix Is Misspecified

Shenghai Dai, Indiana University Bloomington; Dubravka Svetina, Indiana university Bloomington; Cong Chen, University of Illinois at Urbana–Champaign

This study investigates the impact of missing data, and six approaches (i.e., listwise, treating as incorrect, EM imputation, two-way imputation, response function imputation, and logistic regression) for dealing with missingness in implementation of CDMs, especially when the Q-matrix is misspecified. Both item parameter recovery and student classification accuracy were examined.

Electronic Board #12

Evaluate the Impact of Rearranging Multiple-Choice Response Options on Score Comparability

Lin Wang, Educational Testing Service

This study investigated if rearranging multiple-choice items' response options might impact test performance. The items' response options of a base test were rearranged to make three variant tests; the four tests were randomly administered. The findings showed practically insignificant impact on score comparability.

Electronic Board #13

Evaluating a Learning Progression Theory: An Application of Multidimensional Item Response Theory

Duy Pham, University of Massachussets Amherst; Scott Monroe, University of Massachussets Amherst; Malcolm Bauer, Educational Testing Service; Caroline Wylie, Educational Testing Service; Craig Wells, University of Massachussets Amherst

The purpose of this study is to examine a statistical procedure to test a theory of three research-based learning progressions (LPs) for middle-school algebra using a multidimensional item response theory (MIRT) model. The item difficulty and interaction among the constructs of the LPs were analyzed to test the theoretical prediction

Electronic Board #14

Item Order Effects on IRT Number-Correct Score Equating

Juan Chen, National Conference of Bar Examiners; Mark Connally, National Conference of Bar Examiners; Mark Albanese, National Conference of Bar Examiners

This study investigated item order effects under the common-item equating to a calibrated-pool design. Specifically, this study compared item parameter estimates and IRT true-score and IRT observed-score equating results based on data collected from the base form, scrambled form, or a combined data set of both forms, respectively.

Electronic Board #15

Detecting Non-Negligible Model Misfit in IRT

Hwanggyu Lim, University of Massachusetts Amherst; Craig Wells, University of Massachusetts Amherst; Hongyu Diao, University of Massachusetts Amherst

This study introduces a range-H0 approach for assessing the fit of an IRT model in which H0 specifies a tolerable amount of misfit. The advantage of this approach is that rejection of H0 implies the model misfit is non-negligible and items in which the model provides good-enough fit are retained.

Electronic Board #16

Linking Invariance across Subgroups for an English Language Proficiency Assessment

Hanwook Yoo, Educational Testing Service; Venessa Manna, Educational Testing Service; Hyeon-Joo Oh, Educational Testing Service

This study evaluated the invariance of linking functions across subgroups for an English language proficiency test. Unlike previous population invariance studies, this study includes extensive examinee background characteristics such as educational background, language exposure, and previous test experience. The result illustrates how score equity assessment supports test fairness.

Electronic Board #18

Assessing the preservation of equity properties in Mixed-format Test Equating

Raffaela Wolf, Pearson Vue

Equity properties are examined to assess equating results for a Mixed-format assessment. The impact of common-item set composition, group ability and form difficulty differences is investigated for seven equating methods (Tucker, Levine Observed Score, Levine True Score, Frequency Estimation, and Chained Equipercentile, IRT True Score, and IRT Observed Score).

Electronic Board #19

Examinees' Perceptions of the National Physical Therapist Examination Computer-Based Testing Environment

Ellen Donald, Florida Gulf Coast University; Robert Dedrick, University of South Florida

This study assessed examinees' perceptions (n=216, 101 testing centers) of the environment where they took the National Physical Therapy Examination. About 19% perceived that the environment prevented them from performing their best. Results are viewed in the context of the Standards for Educational and Psychological Testing (2014).

Electronic Board #20

Exploring Rubric-Related Multidimensionality in Polytomously-Scored Test Items

Daniel Bolt, University of Wisconsin, Madison; Daniel Adams, University of Wisconsin, Madison

A polytomously-scored test item may measure different dimensions/dimensional composites across rating categories. A multidimensional nominal response model can be used to investigate this possibility, as well as the consequences of ignoring such differences. A real data application is presented in the study of item format effects using TIMSS 2007.

Electronic Board #21

Investigating Response Time on Task Effect in an Early Literacy Assessment

Qinjun Wang, University of Minnesota; Michael Rodriguez, University of Minnesota; Scott McConnell, University of Minnesota; Alisha Wackerle-Hollman, University of Minnesota

This paper describes an application of generalized linear mixed model (GLMM) designed to capitalize on the relationship between response time and task taker performance in a computer-based early literacy assessment and to investigate the conditions that influence the direction and strength of this relationship.

Saturday, April 29, 2017
2:45 PM–4:15 PM, Salon A, Meeting Room Level, Invited Session, G1

Peer Review Under the Every Student Succeeds Act of 2015

Session Chair: Ellen Forte, edCount, LLC
Session Discussants: Marianne Perie, University of Kansas; Melissa Fincher, Georgia Department of Education; Patrick Rooney, US Department of Education; Juan D'Brot, National Center for the Improvement of Educational Assessment; Nathan Dadey, National Center for the Improvement of Educational Assessment

In December, 2015, President Obama signed into law the Every Student Succeeds Act (ESSA) as the replacement of the No Child Left Behind version of the Elementary and Secondary Education Act of 1965. In 2016, a number of states submitted required documentation for peer review even while the Department was developing draft regulation and guidance documents relevant to various aspects of ESSA. The elections of November 2016 further complicated matters with a potentially radical shift in federal oversight of states' public school systems. This session will address the purpose of federal peer review and its manifestation both past and present. Presenters will discuss the current position of the US Department of Education as reflected in regulations, guidance, and other media and place peer review within the larger context of obligations framed in the Standards for Educational and Psychological Testing (AERA/APA/NCME, 2014). The panel includes representatives from state departments of education with extensive peer review experience, experts who have helped to design and implement peer reviews and other forms of technical quality evaluation, and a representative from the US Department of Education who is engaged in the current peer review process.

Saturday, April 29, 2017
2:45 PM–4:15 PM, Conference Room 1&2, Meeting Room Level, Coordinated Session, G2

Connecting Psychometrics and Classroom Assessment to Improve Theory and Practice

Session Chair: Anthony Albano, University of Nebraska-Lincoln
Session Discussant: Michael Rodriguez, University of Minnesota-Twin Cities

Recent changes to educational policy, along with technical and technological advances in the field of educational measurement, encourage the integration of new forms of assessment into classroom instruction and the use of new methods of modeling and validating assessment results to support both formative and summative decision-making. This coordinated session presents five key issues requiring our attention as we consider the future of classroom assessment and its expanding uses within educational and educational accountability systems. These issues include:

- Establishing a theory on the role of classroom assessment in student achievement;
- Using learning theory and validity theory to improve classroom assessment research, design, and implementation;
- Constructing a validity argument for the many purposes of classroom assessments;
- The role of assessment literacy in classroom assessment design and implementation; and
- Measuring classroom assessment practices for the certification of accomplished teachers.

Establishing a Theory of the Role of Classroom Assessment in Student Achievement
Heidi Andrade, University at Albany-SUNY

Using Learning Theory and Validity Theory to Improve Classroom Assessment
Sarah Bonner, Hunter College, CUNY

Constructing a Validity Argument for the Many Purposes of Classroom Assessments
Suzanne Lane, University of Pittsburgh

The Role of Assessment Literacy in Classroom Assessment Design and Implementation
Anthony Albano, University of Nebraska-Lincoln

Measuring Classroom Assessment Practices for the Certification of Accomplished Teachers
Carol Ezzelle, National Board for Professional Teaching Standards

Saturday, April 29, 2017
2:45 PM–4:15 PM, Conference Room 3&4, Meeting Room Level, Coordinated Session, G3

Methodological Advances in PISA

Session Organizer: Leslie Rutkowski, Centre for Educational Measurement at the University of Oslo
Session Discussant: John Mazzeo, ETS

As a means of communicating recent innovations to the broader educational measurement community, this session will bring together recognized scholars in the field of international large-scale assessment to present and discuss the state-of-the-science in the 2015 cycle of the OECD's Programme for International Student Assessment (PISA). Through four connected papers augmented by a discussant, this session will highlight operational solutions for dealing with cross-cultural measurement and comparability, design changes to better estimate trends, procedures for detecting and correcting for response styles, and original uses for process data. Researchers in this session have been conducting inquiry into quantitative methods for two decades or more. Additionally, measurement and assessment specialists who have been directly involved and who have valuable insights about practical issues in international assessments will also participate. Dr. John Mazzeo offers his perspectives as discussant in this coordinated session.

Innovations in Scale Comparability of International Assessments
Matthias von Davier, ETS

Characterization of Linking Error in PISA
Jonathan Weeks, ETS; Matthias von Davier, ETS; Kentaro Yamamoto, ETS

Measuring Response Styles in Rating Data Using Multi-Process IRT Models
Lale Khorramdel, ETS; Artur Pokropek, Instytut Badań Edukacyjnych; Matthias von Davier, ETS

Group-Level Item-Fit Statistics for the Analysis of Invariance
Janine Buchholz, German Institute for International Educational Research; Johannes Hartig, German Institute for International Educational Research

Feature Generation and Selection Using Process Data in Scenario-Based Items in PISA
Qiwei He, ETS; Zhuangzhuang Han, Teacher's College, Columbia University; Matthias von Davier, ETS

Saturday, April 29, 2017
2:45 PM–4:15 PM, Salon B, Meeting Room Level, Coordinated Session, G4

Designing and modeling performance-based assessments of student collaboration

Session Chair: Peter Halpin, New York University
Session Discussant: Patrick Kyllonen, ETS

Collaboration and group work are highly-valued classroom practices, and they have featured prominently in current initiatives concerning the measurement of ``21st century skills." However, fundamental questions remain about how to design and model assessments of collaboration. In particular, while self-reports and situational judgment tasks can be useful research tools, they are not ideal for making consequential decisions about students in educational settings. One way to address this issue is through the use of performance-based assessments, which can provide observable evidence about student competencies. To this end, each paper in this symposium presents novel research on the performance-based assessment of collaboration. The first paper emphasizes design principles based on recent theoretical work on how students learn and solve problems in collaborative settings. The second paper analyses data from three novel task designs using a multi-group IRT-based approach. The third paper studies response time distributions in a set of simulation- and collaboration-based assessments that involve students interacting to solve the same problem. The fourth paper addresses how to simulate data from complex interacting systems, in order to analyze the psychometric properties of models used to study collaboration.

Designing Principles for Collaborative Assessment: Aligning Assessment with How Students Learn
Lei Liu, ETS; Jiangang Hao, ETS; Alina von Davier, ETS; Patrick Kyllonen, ETS

IRT-based models for online tasks that involve student collaboration
Peter Halpin, New York University; Yoav Bergner, New York University

Response-time analysis in simulation-based and collaborative assessments
Lu Ou, Penn State; Jiangang Hao, ETS; Yi-Hsuan Lee, ETS; Lei Liu, ETS; Alina von Davier, ETS

Designing Simulation Studies for Collaborative Tasks: A Psychometric Perspective
Alina von Davier, ETS; Mengxiao Zhu, ETS; Jiangang Hao, ETS; Lu Ou, Penn State

Saturday, April 29, 2017
2:45 PM–4:15 PM, Salon C, Meeting Room Level, Coordinated Session, G5

Issues with Interpreting Measurement Results in Social Sciences: Causality, Dimensionality and Scales

Session Chair: David Torres Irribarra, MIDE, Pontificia Universidad Católica de Chile
Session Discussant: Ronli Diakow, New York City Department of Education

Since their introduction by Spearman (1904), latent variable models have gained widespread acceptance as a tool for measurement in the social sciences among researchers and practitioners alike. Throughout the psychometric literature, various assumptions are made regarding the relationship between the underlying latent variables, and the models, parameters, and indicators being used to attempt to measure them. These assumptions influence the interpretations and use of measurement results in ways that are not always supported by the theoretical frameworks underlying these models. This symposium discusses several issues in the interpretation of measurement results, with specific focus on the relationship between a latent variable and its representations. Theoretical perspectives on this relationship are presented in the first two papers. The first discusses the representation fallacy, which occurs when model parameters are conflated with the attribute being measured, and the second discusses the role that causal interpretations play in understanding the relation between latent variables and their indicators. The second half of the symposium highlights two issues in connecting latent variables to measurement models. The third paper focuses on the use of interval scales in measurement, while the final paper discusses the tension between multidimensional cognitive models and unidimensional measurement models.

Quantitative Attributes, Interval Scales, and the Representational Fallacy
Andrew Maul, University of California, Santa Barbara; Josh McGrane, Oxford University Centre for Educational Assessment

Causal Interpretations and the Interpretation of Causal Mechanisms in Measurement
David Torres Irribarra, MIDE, Pontificia Universidad Católica de Chile; Andrew Maul, University of California, Santa Barbara

Interpretation and considerations in equal difference measurement: Four paradigms for interval scales
Rebecca Freund, University of California, Berkeley

Complex Theories and Simple Models: Expecting Multidimensional Interpretations from Unidimensional Models
Amy Arneson, University of California, Berkeley; David Torres Irribarra, MIDE, Pontificia Universidad Católica de Chile

Saturday, April 29, 2017
2:45 PM–4:15 PM, Salon D, Meeting Room Level, Paper Session, G6

Models for Polytomous Data

Do Option-Based Diagnostic Models Provide More Information than Correct-Option Based Diagnostic Models?
Yanyan Fu, The University of North Carolina at Greensboro; Oksana Naumenko, The University of North Carolina at Greensboro; John Sessoms, The University of North Carolina at Greensboro; Robert Henson, The University of North Carolina at Greensboro; Louis DiBello, University of Illinois at Chicago; William Stout, University of Illinois at Urbana Champaign

The study used simulation and a real dataset to compare the differences between a GDCM-MC and a dichotomous GDCM. The results showed that the GDCM-MC had higher CCRs than the D-GDCM, and the GDCM-MC had better item discrimination than the D-GDCM for both simulation and real data studies.

Exploring polytomous IRT estimation methods using SAS procedures
Sunil Lamsal, Pearson VUE

Various estimation techniques have been developed for the unidimensional polytomous item response theory models. This study focused on comparing the non-linear mixed models procedure and the Markov chain Monte Carlo procedure for estimating partial credit and generalized partial credit models under different sample sizes, test lengths, and response category conditions

Investigating Partial Credit Scoring for Cloze Drop-Down Items
Shuqin Tao, Curriculum Associates; Dan Mix, Curriculum Associates

This study investigates partial credit scoring methods for cloze drop-down, a technology-enhanced item type widely used in CCSS assessments. Data came from a CCSS-aligned standards mastery assessment. Findings will shed light on the reliability and validity aspects of various scoring methods for items with various characteristics.

Using Signal Detection Theory and IRT Methods with Multiple Response Items
Joe Betts, Pearson VUE; William Muntean, Pearson VUE; Doyoung Kim, NCSBN

This research evaluates the analysis of multiple response items using a number of different psychometric models. The models used are the Partial Credit Model, a true-false testlet model, and an item response signal detection model. Unique information related to the different psychometric models will be highlighted.

Saturday, April 29, 2017
2:45 PM–4:15 PM, Salon K, Meeting Room Level, Paper Session, G7

Generalizability Theory

Estimating Domain Scores for Testlets Using Generalizability Theory Approaches

Jiwon Choi, University of Iowa /ACT, Inc.; Xiaohong Gao, ACT, Inc.; Won-Chan Lee, University of Iowa

This study investigates estimating domain scores for testlet-based tests under a generalizability theory framework. Several factors such as ignoring testlet configurations, aggregating item scores, and treating items as nested within testlets are considered. The effect of sample sizes and the number of items within a testlet are also considered.

Investigating the Use of Multivariate Generalizability Theory for Evaluating Subscores

ZHEHAN JIANG, University of Kansas; Mark Raymond, National Board of Medical Examiners

Conventional methods for evaluating the utility of subscores rely on correlations among subscores, but overlook score profile variance. The current study investigates the properties of , a reliability-like index for score profiles based on multivariate generalizability theory, and documents it relationship to traditional indices.

Maximizing Reliability Under Budget Constraints Using G Theory: New Applications, Accessible Approaches

Frank Padellaro, University of Massachusetts; Robert Keller, Measured Progress; Lisa Keller, University of Massachusetts

While Generalizability Theory allows an analyst to estimate and manipulate specific sources of measurement error, an obvious, but practically untaken next step is to use this information to maximize reliability under a budget constraint systematically. This study introduces new software to make the procedure practical for everyday use.

Multifacet Generalizability Designs to Test Reliability

Wei Tao, ACT, Inc.; Yi-Fang Wu, ACT, Inc.

In this study, we utilized several multifacet generalizability theory models to estimate test form reliability, taking into account the complexity of real data structure. Formulas of the appropriate generalizability coefficients were derived, followed by the development of SEM and CSEM. Interval estimates of the generalizability coefficients were also investigated.

Using G-theory with Ordinal Measures to Enhance Accuracy of Reliability Estimation

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

Structural equation modeling techniques are used to extend applications of Generalizability Theory to ordinal-level data. Results highlight the importance of accounting for random-response, specific-factor and transient measurement error and the greater precision obtained when data are properly treated as ordinal rather than equal interval in nature.

Saturday, April 29, 2017
2:45 PM–4:15 PM, Salon L, Meeting Room Level, Paper Session, G8

Technical Issues with Linking & Equating

An Index to Evaluate the Efficacy of Equating Using the Stocking-Lord Method

Unhee Ju, Michigan State University; Louis Roussos, Measured Progress; Lei Yu, Measured Progress

The Stocking-Lord method is widely used in practice. However, little research has conducted on evaluating whether the method is working as intended in linking two scales through common items. This study proposes an index to evaluate its efficacy, and the index is studied with both simulated and real data.

An Iterative Procedure to Detect Item Parameter Drift in Equating Items

Jing Jiang, Boston College; Louis Roussos, Measured Progress; Lei Yu, Measured Progress

Checking the stability of equating items is an important step in test equating. However, little research proposes procedures to detect drifts in pre-equated test items. A new iterative procedure is proposed to detect item parameter drift and select stable equating items to ensure the accuracy of post-equating results.

Comparing Non-Equivalent Groups Linking Methods With an Anchor Test and Covariates

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

This study examines the performance of observed-score linking methods that utilize external covariates to supplement an anchor test in a non-equivalent groups design. Resampling and pseudo-test forms from a state assessment were used to examine linking accuracy in terms of RMSE across different sample sizes and form difficulty differences.

Investigation of Relative Equating Precision for Chained Equipercentile Method

Yanlin Jiang, Educational Testing Service

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

Saturday, April 29, 2017
2:45 PM–4:15 PM, Salon M, Meeting Room Level, Paper Session, G9

Testing Special Populations

Session Chair: Nathan Dadey, Center for Assessment

Case Study: An Instance of Very Strong Local Item Dependence

Han Yi Kim, Measured Progress; Louis Roussos, Measured Progress

After analyzing hundreds of unidimensionally scored tests, we have for the first time found an alternate assessment program where strong multidimensionality occurred for every test. We provide a detailed description of the results, reveal the cause of the multidimensionality, and discuss possible methods for preventing this particular form of multidimensionality.

Evaluating an Initialization Tool for Student Placement into a Map-Based Assessment

Brooke Nash, University of Kansas - Center for Educational Testing and Evaluation; W. Jake Thompson, University of Kansas - Center for Educational Testing and Evaluation

Balancing test length with content targeted towards students' knowledge and skills is especially challenging when assessing students with significant cognitive disabilities. An initialization tool was developed to place students into map-based assessments at levels that match students' skills. This paper presents findings on how well the tool meets this purpose.

Exploring Dimensionality of Data Produced by the NCSC Assessments

Nathan Dadey, Center for Assessment

This study explores the empirical dimensionality of the National Center and State Collaborative assessments. Exploratory factor analytic methods are used to provide groupings of items by dimension, which are then examined by content experts to determine whether the dimensions are construct relevant.

Exploring Teacher Choice When Using an Instructionally Embedded Alternate Assessment System

Amy Clark, University of Kansas; Meagan Karvonen, University of Kansas; Russell Swinburne Romine, University of Kansas; Brooke Nash, University of Kansas

This presentation explores teacher choice during use of an instructionally embedded alternate assessment (AA-AAAS) system. Data from 2015-16 (N = 17,265 students) were reviewed to analyze teachers' choice of extended content standards and levels of complexity for individual student assessments, and the frequency and timing of assessment administration.

Research Pathways Towards Developing an Alternate English Language Proficiency Assessment

Laurene Christensen, University of Wisconsin - Madison; Olivia Lickteig, National Center on Educational Outcomes; Vitaliy Shyyan, National Center on Educational Outcomes

This presentation shares the results of teacher interviews and classroom observations conducted with English learners with significant cognitive disabilities. The study was designed to generate findings about current instructional and assessment strategies supporting the English language development of this population and improve the connection between large-scale assessment and classroom assessment.

Saturday, April 29, 2017
4:35 PM–6:05 PM, Salon A, Meeting Room Level, Invited Session, H1

Psychological and Social Measurement: The Career and Contributions of Benjamin D. Wright

Session Chairs: Mark Wilson, BEAR Center, University of California, Berkeley;
William Fisher, BEAR Center, University of California, Berkeley

Chair's remarks: Book Launch- Mark Wilson

Presenters:

George Engelhard Jr.
Cogitations on invariant measurement: A memo to Ben Wright on the perspectives of Rasch and Guttman

Interlude: Ben Wright in his own words (read by Karen Draney and Stefanie Wind

Mary Lunz & John Stahl
Ben Wright: A multifacet analysis

William Fisher
Provoking Professional Identity Development: The Legacy of Benjamin Drake Wright

Open invitation for members of the audience to share their memories of Ben

Saturday, April 29, 2017
4:35 PM–6:05 PM, Conference Room 1&2, Meeting Room Level, Coordinated Session, H2

Psychometrics for noncognitive assessment: Unique challenges and some solutions

Session Chair: Patrick Kyllonen, Educational Testing Service
Session Discussant: Alberto Maydeu-Olivares, University of South Carolina

Recent research has documented the importance of noncognitive skills in school, work, and life. This has led to a growing call for the use of noncognitive assessment in education including K-12 accountability, monitoring noncognitive skills growth longitudinally, college and graduate school admissions, student progress monitoring, and post-secondary preparedness. Lack of good measurement has been a critical barrier to progress in implementing noncognitive assessment. The dominant measurement methodology has been self-ratings using Likert scales, but these are susceptible to impression management, coaching, faking, and construct-irrelevant response style effects. The purpose of the symposium is to explore alternatives to the simple Likert scale for noncognitive assessment for various purposes ranging from selection to growth and development. The papers identify and address challenges in noncognitive assessment, focusing on skills and attributes ranging from momentary affect to time management, team work, resilience, intrinsic motivation, creativity, and ethics. Assessments include situational judgment tests, repeated measurement of rating scales, and forced-choice methods, and psychometric models include the nominal response model and Bayesian estimation approaches. The measures and psychometric methods presented here may be the most promising approaches today, and research to develop them is critical for future noncognitive assessment.

Cross validation of the NRM-scoring method for a situational judgment test
Hongwen Guo, Educational Testing Service; Jiyun Zu, Educational Testing Service; Patrick Kyllonen, Educational Testing Service

Extreme Response Style and Measurement of Intra-Individual Variability in Affect
Daniel Bolt, University of Wisconsin, Madison; Sien Deng, University of Wisconsin, Madison

Item Analysis Methods for Forced-Choice Assessments
Longjuan Liang, Educational Testing Service

New Developments in Modeling Multi-Unidimensional Forced Choice Data
Jimmy de la Torre, The University of Hong Kong; Iwin Leenen, Instituto Nacional para la Evaluación de la Educación de México; Pedro Hontangas, Universidad de Valencia; Daniel Morillo, Universidad Autónoma de Madrid; Vicente Ponsoda, Universidad Autónoma de Madrid

Saturday, April 29, 2017
4:35 PM–6:05 PM, Conference Room 3&4, Meeting Room Level, Coordinated Session, H3

Innovative Approaches to Fairly Designing and Developing Noncognitive Measures for Diverse Populations

Session Chair: Maria Oliveri, Educational Testing Service
Session Discussant: Kadriye Ercikan, University of British Columbia

Principled design, development, and use of noncognitive assessments that considers the need of culturally and linguistically diverse examinees are needed to support valid score-based inferences for all examinees. The four presentations illustrate novel approaches to conceptualize and develop fair assessments spanning across all test development activities (e.g., from construct conceptualization to score analysis and interpretation). In contrast to typically employed group-based methods based on differential item functioning to establish fairness, the presented approaches examine more deeply the impact of cultural and linguistic diversity on assessment conceptualization, use, and interpretation. The presenters discuss that in an era of increased population diversification, one size does not fit all; instead more carefully crafted approaches may help establish fairness while preserving equal access to cognitive resources for all populations. These presentations seek to augment summative scores derived from assessments with contextual data from examinees to enhance test score interpretation to inform policy and practice meaningfully. They support foundational validity and fairness concepts outlined in the AERA, APA, and NCME (2014) Standards, which place notions of universal design and accessibility to all examinees at the center of test design and development activities in support of valid score-based interpretation.

Effect of Cultural Orientation Differences on Non-Cognitive Factors: Implications for Assessment Design
Edynn Sato, Sato Education Consulting LLC

One Size Does NOT Fit All
Robert Mislevy, Educational Testing Service; Maria Oliveri, Educational Testing Service; Bryan Maddox, University of East Anglia; Rene Lawless, Educational Testing Service

Fairness of Non-Cognitive Measures Depends Upon Their Uses
Kurt Geisinger, Buros Center of Testing

Facets of Fairness in Performance Assessments
Alina von Davier, Educational Testing Service; Jessica Andrews, Educational Testing Service; Thales Ricarte, Educational Testing Service; Jiangang Hao, Educational Testing Service; Lei Liu, Educational Testing Service

Saturday, April 29, 2017
4:35 PM–6:05 PM, Salon B, Meeting Room Level, Coordinated Session, H4

Flexible K-12 Assessments Afforded by ESSA: Psychometric Possibilities and Case Studies

Session Organizer: Denny Way, Pearson
Session Discussant: Laurie Wise, HumRRO

The Every Student Succeeds Act (ESSA) was passed late in 2015 to replace NCLB, which shaped annual statewide assessments in the U.S. for 15 years. Several ESSA provisions provide greater flexibility for states and districts in designing and administering assessments that meet federal guidelines. This symposium explores some of the psychometric possibilities afforded by ESSA and shares experiences from states where innovative assessment pilots and development efforts have been launched. The five papers in the symposium address the range and depth of this topic. The paper topics are:

- Consideration of the psychometric implications of different state assessment models implied by ESSA.
- Design and modeling choices involved in interim assessment systems that produce a single summative score.
- State models being developed for the ESSA "Innovative Assessment and Accountability Demonstration Authority," which waives many aspects of federal regulation.
- An evaluation of the New Hampshire PACE project, a competency-based education model that supports the accumulation of assessment evidence over the year using a mixture of standardized and flexible determinations.
- Statewide innovative science assessments designed to measure the Next Generation Science Standards.

A distinguished discussant will enrich the symposium presentations. Audience dialogue on this important topic will be solicited.

Psychometric Considerations for State Assessment Designs Under ESSA
Denny Way, Pearson; Michael Young, Pearson

Opportunities Afforded by Multiple Assessments: Considering Systems of Interim Assessments
Nathan Dadey, Center for Assessments

Considerations for Maintaining Equity within an Innovative Assessment and Accountability Demonstration Authority
Susan Lyons, Center for Assessment

The Evaluation of the Performance Assessment of Competency Education (PACE) Program
D. E. (Sunny) Becker, HumRRO; Arthur Thacker, HumRRO

Assessing the Next Generation Science Standards: Challenges and Innovative Designs
Brian Gong, Center for Assessment

Saturday, April 29, 2017
4:35 PM–6:05 PM, Salon C, Meeting Room Level, Coordinated Session, H5

Development and Implementation of a Comprehensive Alignment Evaluation Framework

Session Chair: Elizabeth Towles, edCount, LLC

Current alignment evaluation methods yield information that accounts for only a portion of what is considered comprehensive evidence of alignment between an assessment and the academic content and performance standards on which it is based. Forte (2016) builds upon historic and current approaches to alignment by connecting multiple sources of evidence that ultimately provide a more complete alignment evaluation framework. This session begins with a discussion of the alignment evaluation approach described in Forte (2016) and the connections to other approaches used historically. The second presentation transitions to a discussion of the application of Forte's framework across three different evaluations of alignment, and the lessons learned as a result of those evaluation studies. The third presentation focuses on a state's perspective, specifically a state's use of the results from an application of the Forte (2016) alignment evaluation framework. The last presentation reflects on the application of Webb's original alignment model over time and the connections to current alignment perspectives.

Development and Implementation of a Comprehensive Alignment Evaluation Framework
Ellen Forte, edCount, LLC; Lauren Deters, edCount, LLC

Development and Implementation of a Comprehensive Alignment Evaluation Framework
Lori Nebelsick-Gullett, edCount, LLC; Melissa Fincher, Georgia Department of Education; Sara Christopherson, Wisconsin Center for Education Products and Services

Saturday, April 29, 2017
4:35 PM–6:05 PM, Salon D, Meeting Room Level, Paper Session, H6

DIF Session 2

Controlling Type I Error Rates for DIF Assessment with Multiple Grouping Variables
Chi-Chen Chen, National Sun Yat-sen University; Hsiu-Yi Chao, National Chung Cheng University; Jyun-Hong Chen, National Sun Yat-sen University

When grouping variables are correlated, to assess DIF sequentially may cause the inflation of type I error rates. This study suggests a procedure for simultaneous DIF assessment, conducting a simulation study to evaluate its efficacy. Results indicate that the simultaneous assessment can control type I error rate quite well.

Exploration of DIF Grouping Methods: An Empirical Study across Two Countries
Yi-Hsin Chen, University of South Florida

Although DIF could be identified with manifest grouping variables, there may be superior methods for separating examinees to detecting DIF of test items. The study was intended to explore grouping methods for DIF detection and explain the potential sources of DIF items.

Fallacy of Differential Item Functioning Assessment under Cognitive Diagnostic Models
Xue-Lan QIU, The Education University of Hong Kong; Wen-Chung WANG, The Education University of Hong Kong

Existing studies manipulated DIF under the CDM framework as having different slip and/or guessing parameters, but an identical Q-matrix, across groups of examinees. This study clarifies such a fallacy on DIF manipulation and proposes a new method for DIF assessment. Simulation results reveal the superiority of the new method.

General Logistic Latent Growth Models: An Inherently Nonlinear Growth Modeling Framework
Jaehwa Choi, The George Washington University; Ji Hoon Ryoo, University of Virginia; Jeffrey Harring, University of Maryland

In this article, firstly, we propose a new framework of inherently nonlinear latent growth models based on various types of logistic functions (traditional, truncated, and semi-truncated) with a rationale for its use, analytical derivations, and theoretical/practical implications as a growth modeling technique

Predicting Response Time on Pretest Items from Item Features
Stephen Cubbellotti, American Board of Internal Medicine; Bozhidar Bashkov, American Board of Internal Medicine

In this study, we built and examined the effectiveness of a model predicting item response time on pretest items using item features (e.g., media) in three different exams. In some cases, adding item features to the model improved prediction accuracy over a word-count-only model. Implications for test construction are discussed.

Saturday, April 29, 2017
4:35 PM–6:05 PM, Salon K, Meeting Room Level, Paper Session, H7

Issues with Computerized Assessments

Biased and Unbiased Test Reliability Estimators for Multistage Computerized Adaptive Testing
Yi-Fang Wu, ACT, Inc.

This study investigates biased and unbiased test reliability estimators of multistage computerized adaptive tests using a 3PL IRT model. Furthermore, two alternatives of test information function (Samejima, 1990) are utilized to enhance reliability measures in terms of improving measurement precision for individuals of lower and higher ends of ability levels.

Comparing Computer-Adaptive Testing Stopping Rules Using the Generalized Partial Credit Model
Christopher Runyon, The University of Texas at Austin; Rose Stafford, The University of Texas at Austin; Jodi Casabianca, The University of Texas at Austin; Barbara Dodd, The University of Texas at Austin

Various stopping rules for computer-adaptive tests exist, however their relative performance under the generalized partial credit model has not been studied. This simulation study compares eleven stopping rule conditions with item exposure controls and content balancing. The standard error stopping rule provides accurate theta estimates while maintaining operational efficiency.

Evaluating CAT item pool feasibility: Investigating two software programs
Ryan Wilke, Florida State University; Cody Diefenthaler, Florida State University; Richard Luecht, University of North Carolina, Greensboro

One challenge confronted in computerized adaptive testing (CAT) is creating adequate item pools. We investigated two software programs that assess CAT item pool feasibility; that is, where on the theta scale CAT criterion can be achieved for an item pool, given item characteristics, stopping rules, and item exposure requirements.

Investigation of Psychometric Properties of Technology-Enhanced Items Considering Content Characteristics
Rong Jin, Houghton Mifflin Harcourt Publishing; Stephen Murphy, Houghton Mifflin Harcourt Publishing; Sid Sharairi, Houghton Mifflin Harcourt Publishing

Recent research explored technology-enhanced (TE) items only by subject. This study dives into DOK, master domain, and grade level to explore over six hundreds TEs in twelve formats from a K-8 large scale mathematics assessment and compares the measurement properties between multiple-choice and TE, and between TE formats.

Item Parameter Drift Analysis for a Computer Adaptive Test
Yufeng Chang, Minnesota Department of Education; Changjiang Wang, Pearson; Bethany Bynum, HumRRO; Gerald Griph, Pearson; Kevin Cappaert, Pearson

We propose a combination of statistical, visual, and judgmental procedures for drift analysis for a CAT program. With the suggested procedures and principles, we hope to provide guidance for assessment practitioners to safeguard the healthiness as well as sustainability of the CAT programs.

Saturday, April 29, 2017
4:35 PM–6:05 PM, Salon L, Meeting Room Level, Paper Session, H8

Reliability and Validity of Classroom Assessment

A Perfect Storm for Progressive Assessment Practice: The Educational Maker Movement

Yoav Bergner, New York University; Samuel Abramovich, University at Buffalo

This paper examines the educational Maker Movement as a challenge/opportunity for modern assessment. Historical-philosophical roots are considered as are current claims about learning and growth in Makerspaces and Fablabs. Assessment frameworks suitable for substantiating these claims are discussed in the context of bridging ideology and practice in progressive education.

Predictive Utility of Teacher Practice Assessment and Growth in Achievement

Adam Lekwa, Rutgers, the State University of New Jersey; Christopher Dudek, Rutgers, the State University of New Jersey; Ilona Arnold-Berkovit, Rutgers, the State University of New Jersey; Jiefang Hu, Rutgers, the State University of New Jersey; Linda Reddy, Rutgers, the State University of New Jersey; Ryan Kettler, Rutgers, the State University of New Jersey

The predictive validity of the Classroom Strategies Assessment System, an assessment of teacher practices, for student growth in achievement was evaluated. Results indicated a significant relationship between teacher practices and gains in achievement. Implications for future practice and research will be discussed.

Promoting In-Depth Examination of Teacher-Developed Placement Tests

Deni Basaraba, Bethel School District #52; Pooja Shivraj, Research in Mathematics Education Southern Methodist University

Educators and researchers alike voice concerns about the technical adequacy of teacher-developed assessments. The purpose of this study is to examine the technical rigor and classification accuracy of a Grade 8 science placement test, with the intent of compiling practically and instructionally meaningful evidence for educators.

Psychometric Analysis of a Statewide Kindergarten Electronic Portfolio Assessment

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

This study provides preliminary evidence for the validity and reliability of a Statewide Kindergarten Electronic Portfolio Assessment in 2015, which was the first effective year of implementation. Classical Test Theory and Item Response Theory analyses were performed with a sample of 113 districts, 1,105 schools, 4,389 teachers and 85,339 kindergartners.

Teacher Practice Data and Interim Student Assessment in Predicting Statewide Performance

Christopher Dudek, Rutgers, the State University of New Jersey; Adam Lekwa, Rutgers, the State University of New Jersey; Linda Reddy, Rutgers, the State University of New Jersey

Data from approximately 400 teachers and 3000 students in grades 3 through 8 were analyzed to evaluate the extent to which data on teacher practices could impact prediction of state test outcome by interim assessment in reading and math. Results and implications for practice and future research will be discussed.

Saturday, April 29, 2017
4:35 PM–6:05 PM, Salon M, Meeting Room Level, Paper Session, H9

Advances in Multilevel Modeling

An Ordinal Approach to Decomposing Test Score Gaps

David Quinn, University of Southern California; Andrew Ho, Harvard Graduate School of Education

We develop and evaluate methods for decomposing ordinal test score gaps into within- and between-school portions. Unlike existing parametric decomposition methods, our methods have the advantage of being scale-invariant across monotonic scale transformations.

Choosing Mathematical Functions for Individual Academic Growth

Gary Williamson, MetaMetrics

Choosing a mathematical function for individual growth has implications for the validity of test-based accountability metrics. The interpretation and quantification of individual growth can be improved by systematically considering: (a) features of growth, (b) mathematical characteristics of prospective growth functions, (c) practical utility, and (d) conjoint measurement.

Measuring Intra-Individual Change on More Than Two Occasions with Hypothesis Testing Methods

Chaitali Phadke, University of Minnesota; David Weiss, University of Minnesota

We propose omnibus hypothesis testing methods for detecting psychometrically significant change at an individual level when an individual is measured on more than two occasions. We also investigate their Type I error and power in the context of the adaptive measurement of change (AMC) and apply these to K-12 data.

NONPARAMETRIC MULTILEVEL LATENT CLASS ANALYSIS: AN APPROACH TO CLASSIFICATION IN MULTILEVEL CONTEXTS

Chi Chang, Michigan State University

The research aims to investigate how classification accuracy of a nonparametric multilevel latent class analysis with covariates are affected by study factors: 1) the CRP of indicators (i.e. Skills), 2) the number of indicators, 3) covariate effects at different levels 4) sample size. A total of 384 conditions were examined.

Saturday, April 29, 2017
6:30 PM–8:00 PM, Rio Vista Room, Meeting Room Level

NCME President's Reception

By invitation Only.

Sunday, April 30, 2017
8:15 AM–10:15 AM, Salon A, Meeting Room Level, Invited Session, I1

Assessing Student Learning Outcomes in Higher Education

Session Chair: Hamish Coates, Centre for the Study of Higher Education, University of Melbourne

The assessment of students' learning outcomes is a pressing change frontier for higher education in many countries and higher education institutions. This session will be structured to present formative and influential work being done to chart cogent technical foundations for the field of higher education learning outcomes assessment. The presenters will articulate the contributions and complexities of existing development, link this research field with other technical and practical considerations, and build a community for advancing further applied and scholarly work

Best practices, success and challenges of skills assessment in Canadian postsecondary institutions
Fiona Deller; Alexandra MacFarlane; Sarah Brumwell; Elyse Watkins

The effects of faculty engagement in research on student learning
Prashant Loyalka; Zhaolei Shi; Guirong Li; Elena Kardanova; Igor Chirikov; Lydia Liu; Jinghuan Shi; Shangfeng Huff; Ninging Yu; Changqing Xu; Liping Ma; Sophie Guo; Huan Wang

Assessing college critical thinking: Results from the Chinese HEIghten™ critical thinking assessment
Lydia Liu; Prashant Loyalka; Amy Shaw; Lin Guo

Assessing higher education learning outcomes: Advances and challenges in the KoKoHs program
Olga Troitschanskaia; Hans Anand Pant; Susanne Schmidt

Measuring and comparing achievements of learning outcomes in higher education in Europe
Robert Wagenaar

The methodological challenges of measuring student learning, degree attainment, and early labour market outcomes in higher education
Tatiana Melguizo; Gema Zamarro; Tatiana Velazco; Fabio Jose Sanchez

An architecture for governing higher education student learning outcomes assessment
Hamish Coates, Centre for the Study of Higher Education, University of Melbourne

Sunday, April 30, 2017
8:15 AM–10:15 AM, Conference Room 1&2, Meeting Room Level, Coordinated Session, I2

New Development and Applications of Automated Scoring in Educational Assessments

Session Organizer: Jing Chen, HumRRO

Innovative educational assessments tremendously increase scoring demand and pose scoring challenges. To provide feedback quickly and make scoring of the new educational assessments affordable, these assessments will need to rely heavily on automated scoring. Automated scoring has been wildly applied to score responses such as essays, spoken responses, and mathematics responses. Nonetheless, automated scoring has not been wildly applied to other types of responses from recent innovative assessments. This symposium is organized around the theme of developing automated scoring for new types of educational assessment data that automated scoring techniques are barely available. The four proposals included in this symposium report work on inventing new automated scoring techniques in the following four aspects: 1) develop an automated annotation system for conversational data (i.e., text chat) in collaborate problem solving tasks, 2) build automated scoring techniques to evaluate examinees' writing process data (i.e., keystroke logs), 3) investigate an automated classification system to classify science short responses into learning progression levels, and 4) develop automated predictions of item difficulty for reading comprehension items. Research and development of automated scoring techniques for new types of educational assessment data is indispensable for improving measurements, increasing efficiency and lowering scoring cost.

CPS-rater: Automated Sequential Annotation of Collaborative Problem Solving Skills
Jiangang Hao, ETS

Defining Bursts in Writing Processes
Mo Zhang, ETS

Automated Classification of Students' Short Responses into Learning Progression Levels
Jing Chen, HumRRO

Item difficulty estimation by means of Generalized Boosting Model: IDM by GBM
Daniel McCaffrey, ETS

Sunday, April 30, 2017
8:15 AM–10:15 AM, Conference Room 3&4, Meeting Room Level, Coordinated Session, I3

Applications of Regression Concepts to Test Linking

Session Organizer: Tim Moses, College Board

Several concepts from regression methodology have been applied to the psychometric evaluations of items, tests and test linkages. The purpose of this proposed symposium is to present a set of studies expanding how regression concepts are applied to evaluate items, test scores and test score linkings. The first study considers rationales and applications of regression-based projection methods for linking PSAT scales to PISA. The second and third studies demonstrate how the quality of subscore equatings and the implications of using concorded scores in regression equatings can be considered with respect to prediction error. The fourth and fifth studies are concerned with subgroup invariance of item scores regressed on ability (i.e., DIF) and concordances. Each of the presentations will cover existing methods as well as proposals of new methods and new applications of existing methods. An expert in the psychometrics field will provide a final concluding discussion of the presentations.

Applications of Post Stratification Procedures from Test Equating to Norms Estimation
YoungKoung Kim, College Board

Evaluating Equated Subscores With Respect to Prediction Error
Neil Dorans, Educational Testing Service

Prediction Error Implications of Using Concorded Scores in Regression Functions
Tim Moses, College Board

Regression on Observed, True and Scaled Matching Variables in Differential Item Functioning
Anita Rawls, College Board

Assessing Subgroup Intercept Differences in Linear Scaling Functions for Concordance Studies
Pamela Kaliski, College Board

Sunday, April 30, 2017
8:15 AM–10:15 AM, Salon B, Meeting Room Level, Coordinated Session, I4

Issues and Challenges in the Measurement of Noncognitive Skills in Applied Settings

Session Organizer: Carol Barry, College Board
Session Discussant: Jack Buckley, College Board

There has been a recent and growing interest among educators, policy makers, and researchers in understanding and measuring students' noncognitive attributes. Noncognitive attributes can be broadly defined as including behaviors, attitudes, beliefs, and strategies important for student achievement but not included in traditional measures of cognitive ability. The study of these dimensions often includes motivation, self-regulation, personality, and other character strengths. As the focus on these attributes has increased so too has the focus on their measurement. This coordinated session will explore issues and challenges in the applied measurement of noncognitive areas. The four papers in this session will highlight current work being done by testing organizations in a variety of contexts: serving different populations, measuring a range of noncognitive constructs, and using that information for a variety of purposes. Subsequently, each presentation will discuss measurement challenges specific to those contexts as well as methods used to address those challenges and/or the implications for test score use.

Development of Short Questionnaire Scales for the National Assessment of Educational Progress
Jonas Bertling, Educational Testing Service

Addressing Challenges in Measuring Noncognitive Skills with the Tessera Noncognitive Assessment System
Jeremy Burrus, Center for Innovative Assessments, ProExam; Kate Walton, St. John's University; Gabriel Olaru, Center for Innovative Assessments, ProExam; Richard Roberts, Center for Innovative Assessments, ProExam

Examining Assumptions Underlying the Measurement of Noncognitive Attributes
Carol Barry, College Board; Rory Lazowski, College Board; Haifa Matos-Elefonte, College Board

Noncognitive Assessment: Users, Use, and a Theory of Action
Ross Markle, Educational Testing Service

Sunday, April 30, 2017
8:15 AM–10:15 AM, Salon C, Meeting Room Level, Coordinated Session, I5

Putting the Content Back into Validity Evidence

Session Organizer: Catherine Welch, University of Iowa
Session Chair: Steve Dunbar, University of Iowa
Session Discussant: Greg Cizek, University of North Carolina

Whether in formative or summative assessments, scores are used to make inferences about what test takers know and can do. Validity affects the inferences we are able to make from these test scores and content validity is one type of validity that allows us to make claims about what a test measures. Educational assessments attempt to reason from specific things students do, make, or say, to broader inferences about their knowledge and abilities. Without evidence related to content validity, we cannot have confidence in these inferences. The procedures used to develop and revise test materials are the foundation for the assessment's content validity. Meaningful evidence related to inferences based on content and performance standards guides the design and development of the content of any assessment. This session addresses key aspects of content validation that can be used as a basis for a validity argument. Two papers examine validation procedures used for new summative and formative assessments. One paper examines the role of content sampling in the validation process. The final paper suggests a need to more carefully articulate and implement content validity studies in an effort to evaluate assessments.

The essential role of content validity in refining assessments
Sherral Miller, College Board

Necessary evidence for content validity
Angelica Rankin, Project Lead the Way

Content sampling as validity evidence
Ashliegh Crabtree, University of Iowa

Moving the dialogue forward on content validity
Catherine Welch, University of Iowa

Sunday, April 30, 2017
8:15 AM–10:15 AM, Salon D, Meeting Room Level, Paper Session, I6

Test Security Issues

Session Chair: Dennis Maynes, Caveon, LLC

A Method for Estimating the Latent Source Used by Answer Copiers

Dennis Maynes, Caveon, LLC

Sometimes educators disclose the exam's answer key (known as the latent source) to students, a violation of test security. This research shows when the latent source contains errors, it can be estimated and used to compute the probability of answer key disclosure to students and its incidence using Bayesian techniques.

A New Statistic to Detect Potentially Fraudulent Erasures

Sandip Sinharay, ETS; Minh Duong, Pacific Metrics; Scott Wood, Pacific Metrics Corporation

We suggest a new statistic to detect potentially fraudulent erasures at the examinee level. The statistic is based on the likelihood ratio test. The statistic is shown to have satisfactory Type I error rate and power for simulated data and to provide useful information for data from an operational test

Alternating Minimization Approach to Detect Group Cheating

Dmitry Belov, Law School Admission Council

Group cheating is hard to detect due to multiple unknowns: unknown subgroups of examinees have an unfair advantage on unknown subsets of administered items. This proposal presents an algorithm to detect group cheating based on alternating minimization. Its performance is studied on simulated and real data.

Compromised Item Detection for Computerized Adaptive Testing Using Generalized Linear Model

Cheng Liu, University of Notre Dame; Jun Li, University of Notre Dame

Item leakage has been a continuous concern in the large-scale computerized testing, especially computerized adaptive testing (CAT). In this paper, authors develop a more generalized detection model with a leaking parameter to represent probability change dynamics. The new model overall achieves high detection accuracy and low type I error.

Detecting Compromised Items Using Information from Secure Items

Xi Wang, Measured Progress; Yang Liu, University of California, Merced

Two methods are proposed to detect compromised items in a continuous testing program. Examinees' responses on items being administered infrequently are used as a reference to detect compromise on items with high exposure rates. Both methods do not require item-recalibration and can be applied in relatively small sample sizes.

Detecting Test Collusion Using Network Analysis

Jiyoon Park, Federation of State Boards of Physical Therapy; Yu Zhang, Federation of State Boards of Physical Therapy; Lorin Mueller, Federation of State Boards of Physical Therapy

Test collusion is one of the types that commonly occur in test administrations, especially in high stakes tests. In this study, we propose a method for detecting test takers as cluster who potentially colluded during test administrations using network analysis.

Sunday, April 30, 2017
8:15 AM–10:15 AM, Salon K, Meeting Room Level, Paper Session, I7

Multidimensional IRT

A Fused Latent and Graphical Model for Multivariate Binary Data

Yunxiao Chen, Emory University; Xiaoou Li, University of Minnesota; Jingchen Liu, Columbia University; Zhiliang Ying, Columbia University

We propose a new model for item response data. It adds a sparse graphical model component to the multidimensional item response theory model, capturing the local dependence structure. Model selection and parameter estimation are discussed. The model implies a better scoring rule and can be used to improve test design.

Bootstrap-Calibrated Interval/Set Estimation for Multidimensional IRT Scale Scores

Yang Liu, University of California, Merced; Ji Seung Yang, University of Maryland, College Park

In applications of multidimensional item response models, the uncertainty due to the estimation of model parameters is often non-negligible and must be accounted for when calculating latent trait scores. We propose bootstrap-calibrated interval/set estimators for scale scores and evaluate their performance via Monte Carlo simulations.

Estimating Latent Variable Interactions with Non-normal and Likert-type Variables

Ezgi Ayturk, Fordham University; Heining Cham, Fordham University

A Monte-Carlo simulation study is conducted to test the performance of unconstrained product indicator approach to latent variable interactions with continuous and categorical indicators under normal and non-normal conditions. Results showed non-normality, and ordinal variables impaired the estimation performance. Implications in substantive research, and illustrative data example will be discussed.

Item Parameter Recovery for Simple-Structure Multidimensional Model using MCMC Method

Laurentius Susadya, University of Iowa; Stella Kim, University of Iowa; Ye Ma, University of Iowa

Simple-structure multidimensional model (SS-MIRT) as a more complex but useful model hasn't been explored enough in previous research. Current study aims to evaluate the performance of two estimation methods in item parameter recovery for SS-MIRT: Marginal Maximum Likelihood (MML) and Markov Chain Monte Carlo (MCMC).

Item Response Theory Modeling of Cross-Classified and Trifactor Item Cluster Structures

Ken Fujimoto, Loyola University Chicago

This study investigates the consequences of specifying a cross-classified item cluster structure in an item response theory model when a trifactor structure is more appropriate. Of particular interest is the affect this incorrect specification has on measurement reliability and inferences about the sources that induce items to cluster.

The Effect of Non-normal Latent Distributions on Estimating Multi-unidimensional Graded Response Models

Tzu-Chun Kuo, Southern Illinois University Carbondale; Yanyan Sheng, Southern Illinois University Carbondale

Two marginal maximum likelihood, three fully Bayesian, and Metropolis-Hastings Robbins-Monro algorithms were compared for estimating multi-unidimensional graded response models with normal and non-normal latent traits. Preliminary results suggested that the Hastings-within-Gibbs estimation had an overall better parameter recovery than the other methods regardless of the shape of latent traits distributions.

Sunday, April 30, 2017
8:15 AM–10:15 AM, Salon L, Meeting Room Level, Paper Session, I8

Bayesian Approaches to IRT

Session Chair: Aki Kamata, Southern Methodist University

A Multidimensional Confirmatory Model for Item Discriminations

Menglin Xu, The Ohio State University; Paul De Boeck, The Ohio State University

We propose a multidimensional confirmatory model for item discriminations, with a linear predictor for the discriminations and a random residual per dimension. The model is applied to PROMIS data on depression and anxiety, using jags for the estimation. The model performance is compared with conventional CFA and bi-factor models.

A Psychometric Analysis of the Tripod Student Perceptions Survey

Megan Kuhfeld, University of Texas at Austin

This study develops a validity argument for using the Tripod student perception survey to measure teacher practice in summative teacher evaluations. Using data from six school districts, I utilize multilevel item factor analysis models to examine the psychometric properties and validity of inferences about teachers from student surveys.

Bayesian Modeling of Multilevel Item Response Model with Heterogeneous Within-Group Variance

Yusuf Kara, Anadolu University; Akihito Kamata, Southern Methodist University

Multilevel item response model was extended to relax the assumption of variance homogeneity by modeling unique within-group ability variances, adopting the Bayesian approach. A simulation study was conducted under varying degrees of heterogeneity. Simulation results showed that the proposed model performs well, regardless of the within-group heterogeneity degree.

Comparison of standard error estimation procedures: when the model is misspeicified

Xiaoxiao Liu, School of Psychology, Beijing Normal University; Yanlou Liu, China Academy of Big Data for Education,Qufu Normal University; Tao Xin, Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University

The present study compared the performance of the 6 types of standard error estimation procedures included in flexMIRT (Cai, 2015) across 8 model-data specification conditions designed by Yuan, K.-H., Y. Cheng and J. Patton (2014) with simulated data. Sandwich-type covariance matrix and SEM are recommended as routine choices.

The Effect of Recoding on Mixture IRT Models

Youn-Jeng Choi, University of Georgia

This purpose of study is to detect the effect of different recoding on the mixture IRT models. Mixture Generalized Partial Credit IRT Model and mixture 2PL IRT models using two differently recoded responses will be applied.

The impact of Markov chain convergence on estimation of mixture IRT models

Yoonsun Jang, University of Georgia; Allan Cohen, University of Georgia

A non-converged Markov chain can potentially lead to invalid inferences about model parameters. This study investigates the effect of degrees of convergence of an MCMC chain, based on various convergence diagnostics, on accuracy of parameter estimates in mixture IRT models.

Sunday, April 30, 2017
8:15 AM–10:15 AM, Salon M, Meeting Room Level, Paper Session, I9

Diagnostic Classification Models

Session Chair: Shiyu Wang, University of Georgia

Assessing Intervention Effects in a Diagnostic Classification Model Framework

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

The transition diagnostic classification model (TDCM) was recently developed to assess growth in a diagnostic classification model framework. This study extends the TDCM to accommodate multiple groups, thereby enabling the estimation of group-differential growth and the evaluation of intervention effects. It is applied to evaluate an instructional intervention in mathematics.

Incorporating Covariates in Log-Linear Cognitive Diagnosis Model

Manqian Liao, University of Maryland College Park; Hong Jiao, University of Maryland

This study aims at exploring the impact of incorporating covariates into both compensatory and non-compensatory Log-linear DCMs on classification accuracy. It is expected that the DCMs with covariates will improve model parameter recovery and classification accuracy, thus providing more accurate classification decisions about students' skill mastery states.

Measuring Adaptive Learning Progress with Cognitive Diagnostic-Computerized Adaptive Testing

Susu Zhang, University of Illinois at Urbana-Champaign; Haiyan Lin, ACT Inc.; Xiaohong Gao, ACT Inc.; Hua-Hua Chang, University of Illinois at Urbana-Champaign

New cognitive diagnostic-computerized adaptive testing (CD-CAT) algorithms are proposed for post-learning assessments in adaptive learning. By focusing on item selection and estimation of the previously learned attribute and its prerequisites, the new procedure can accurately estimate learners' progress with few items, providing valuable information for content recommendation without interrupting learning.

Tracking skill change using predictors: A model for learning in cognitive diagnosis

Shiyu Wang, University of Georgia; Yan Yang, University of Illinois at Urbana-Champaign; Steven Culpepper, University of Illinois at Urbana-Champaign; Jeff Douglas, University of Illinois at Urbana-Champaign

This study proposes a learning model which integrates a cognitive diagnostic model with a hidden markov model by using covariates to model transition probability. The model is applied in a computer-based assessment with training tools to track the change of students' skills and to evaluate the efficacy of learning interventions.

Two-stage estimation of Bayesian network and diagnostic classification model structural parameters

Jonathan Templin, University of Kansas; Lu Qin, University of Kansas

In this paper, we investigate a method for deriving BayesNets conditional probabilities between latent nodes from separate calibration analyses. The two-stage estimation method uses separately calibrated examinee estimates in a weighted-variable analysis. Via simulation study and empirical data analysis, we show the effectiveness of the method for estimating BayesNets parameters.

Using Diagnostic Classification Models to Detect Model Misspecifications in Bayesian Networks

Bo Hu, The University of Kansas; Jonathan Templin, The University of Kansas

The purpose of this study is to investigate the use of diagnostic classification models in order to provide evidence for model fit for latent variable edges in Bayesian inference networks. Through simulation study and empirical data analysis, we show how the use of nested model comparisons can validate BayesNets models.

Sunday, April 30, 2017
10:35 AM–12:05 PM, Salon A, Meeting Room Level, Invited Session, J1

Fairness in Educational Assessment and Measurement

Session Chair: Linda Cook
Session Discussant: Michael Walker, The College Board

The goal of the NCME edited volume, Fairness in Educational Assessment and Measurement, is to provide information about fairness as it pertains to a wide set of educational topics and issues. . The 16 chapters included in the volume focus primarily on educational assessment in the United States The volume has three parts, plus two introductory and closing chapters written by the editors. Part I describes existing practices and procedures for ensuring fairness in test design, construction, administration and scoring. Part II addresses complexities associated with assessing the fairness of comparisons made under divergent measurement conditions such as tests built to different specifications, or tests administered in different modes of assessment, or across different grades in school, or in different languages, or to special populations. Part III includes chapters that examine fair assessment from different perspectives. Each part includes a commentary. This symposium has four presentations. The first presentation provides an overview of the book. The next three presentations pertain to three chapters selected from Part II. The symposium includes a discussant who has extensive experience in fairness and educational assessment.

Overview
Neil Dorans, Educational Testing Service

The Fairness of Comparing Test Scores Across Different Tests or Modes of Administration –
Mary Pommerich, Defense Personnel Assessment Center

Comparing Test Scores across Grade Levels
Michael Kolen, University of Iowa

Comparing Scores from Tests Administered in Different Languages
Stephen Sireci, University of Massachusetts

Sunday, April 30, 2017
10:35 AM–12:05 PM, Conference Room 1&2, Meeting Room Level, Invited Session, J2

Measuring Creativity from Classrooms to Large Scale Assessments: Views from Practice to Research and Development of Assessments

Session Chair: Bonnie Strykowski, National Association of Assessment Directors and Mesa Public Schools, Arizona

This invited symposium explores issues from classrooms to national and international assessments as the panelists discuss one of the trending topics in international education, economics, and business: creativity. As we begin to experience more programs in our classrooms promoting creativity, many challenges exist in measuring and assessing the results of these endeavors. This session will include an interactive component during which panelists and the audience will discuss relevant issues related to this topic, including the implications for large-scale assessment.

An International Perspective on Creativity and Critical Thinking Skills: Creating Bridges from Classrooms to International Assessments
Stéphan Vincent-Lancrin, OECD Directorate for Education and Skills

A Window into the World of Teaching Creativity in Classrooms
Daniel Wallace, Senior Program Manager, LIncoln Center Education, New York

Toward Valid Measures of Creativity: Challenges and Promising Approaches
Joan Herman, UCLA-CRESST

Innovation and Creativity in the National Assessment of Educational Progress (NAEP)
Sharyn Rosenberg, National Assessment Governing Board

Sunday, April 30, 2017
10:35 AM–12:05 PM, Conference Room 3&4, Meeting Room Level, Coordinated Session, J3

Challenges and Opportunities in the Application of Diagnostic Measurement

Session Chair: Shenghai Dai, Indiana University Bloomington
Session Discussant: Robert Henson, The University of North Carolina at Greensboro

Interest in diagnostic classification model (DCM) has increased among educational researchers and practitioners in the past two decades. Along with the rapid development of DCMs, concerns and challenges arise. The four studies in this session addressed several important concerns in the application of DCMs, including the development of a new test versus the challenges to retrofit DCMs on existing assessments, how to design DCMs that take into account both item-attribute specifications and misconceptions, and how to make multilevel diagnostic inferences in large scale assessment settings.

Retrofitting v.s. Developing Assessments for Diagnostic Purposes: Two Lessons Learned
Ren Liu, University of Florida; Anne Huggins-Manley, University of Florida; Oksan Bulut, University of Alberta

Comparison of Three Q-matrix Validation Methods for Developing Cognitive Diagnostic Assessments
Cong Chen, University of Illinois at Urbana–Champaign; Shenghai Dai, Indiana University Bloomington; Jinming Zhang, University of Illinois at Urbana–Champaign

Option-based and Global Fit Statistics for the GDCM-MC
Yanyan Fu, The University of North Carolina at Greensboro

Making Multilevel Diagnostic Inferences for Large-Scale Assessments
Shenghai Dai, Indiana University Bloomington; Dubravka Svetina, Indiana University Bloomington; Xiaolin Wang, Indiana University Bloomington

Sunday, April 30, 2017
10:35 AM–12:05 PM, Salon B, Meeting Room Level, Coordinated Session, J4

Understanding Student Decision Making using Markov Decision Processes

Session Chair: Michelle LaMar, Educational Testing Service
Session Discussant: Kristen DiCerbo, Pearson

There has long been a tension between the types of performance tasks that educators value for their realistic complexity and the types of assessment items that psychometricians assert are needed to produce reliable and valid measurement. With the availability of increasingly fine-grained data documenting student performances in both learning and assessment tasks, researchers now have the opportunity to re-conceptualize what can be learned about students based on such tasks, and the methods by which to do so. This session explains and demonstrates a new approach to measurement of student performance in complex tasks using the Markov decision process (MDP) to represent student cognition. The MDP provides a framework in which different measurement models can be formulated targeting such constructs as strategic problem solving, domain misconceptions, motivation, and goals. In this symposium we present five different applications of the model to educational assessment based on diverse assessment formats including simulation-based science experiments, algebraic equation solving, interactive cultural simulation games, and spoken dialog with an automated agent. The presentations demonstrate the flexibility and potential of the approach while discussing its limitations and work yet to be done.

An Introduction to the Markov decision process measurement framework
Michelle LaMar, Educational Testing Service; Janet Koster van Groos, Educational Testing Service

Exploring the Detection of a Proportional Reasoning Strategy in a Science Simulation
Rajendra Chattergoon, University of Colorado, Boulder; Michelle LaMar, Educational Testing Service; Gabrielle Cayton-Hodges, Educational Testing Service; Maddy Keehner, Educational Testing Service; Janet Koster van Groos, Educational Testing Service

Understanding Equation Solving Skills Using Inverse Planning
Anna Rafferty, Carleton College; Rachel Jansen, University of California, Berkeley; Tom Griffiths, University of California, Berkeley

Building a Cognitive Model of Implicit Beliefs in Cross-Cultural Competency
LaTasha Holden, Princeton; Malcolm Bauer, Educational Testing Service; Michelle LaMar, Educational Testing Service

Conversational competence in dialogic assessment
Vikram Ramanarayanan, Educational Testing Service; Michelle LaMar, Educational Testing Service

Sunday, April 30, 2017
10:35 AM–12:05 PM, Salon C, Meeting Room Level, Coordinated Session, J5

Understanding the difference between professional and legal expectations: A practitioner's challenge

Session Chair: Chad Buckendahl, ACS Ventures, LLC
Session Discussant: Doug Becker, Houghton Mifflin Harcourt

The Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014) is intended to serve as an important resource for programs and practitioners during test development and validation. However, users who rely primarily on these expectations may be surprised to learn that in legal proceedings, the courts may not interpret or apply the Standards as recommended by the profession. This creates a conundrum for practitioners and policymakers when designing programs and prioritizing validation efforts. In this session, presenters will discuss multiple perspectives that emerged from court proceedings related to a credentialing examination that was interpreted by the courts as an employment test. An initial perspective will discuss a brief history of the case along with the implications for a range of stakeholders impacted by this decision. A second perspective will discuss the experiences and lessons learned from the presenter's participation as an expert witness during the proceedings. A third perspective will represent a policymakers' challenge of responding to a legal decision for an analogous program in another jurisdiction while considering professional guidelines. Finally, a psychometric practitioner's perspective about how to balance legal requirements with the expectations of the Standards with recommendations for practice.

Understanding Gulino: A legal perspective
Michelle Croft, ACT, Inc.

Applying the Standards to Gulino: An expert witness perspective
Chad Buckendahl, ACS Ventures, LLC

Responding to Gulino: A regulatory agency perspective
Phil Canto, Florida Department of Education; Lauren White, Florida Department of Education

Responding to Gulino: A psychometric practitioner perspective
Susan Davis-Becker, ACS Ventures, LLC; Renee Launey-Rodolf, Oklahoma Office of Educational Quality and Accountability

Sunday, April 30, 2017
10:35 AM–12:05 PM, Salon D, Meeting Room Level, Paper Session, J6

Vertical Scaling

Investigation of construct shift in vertical scaling using multidimensional IRT model
Mingcai Zhang, Michigan State University; Lihong Yang, Australian Curriculum, Assessment and Reporting Authority

To evaluate the impact of construct-shift in vertical scaling, the unidimensional and multidimensional IRT models are compared in developing a single vertical scale across grades. Simulation conditions include construct correlation, average achievement growth, and anchor item design. The reference composite based on MIRT approach performs better in improving score comparability.

Investigation of construct shift in vertical scaling using multidimensional IRT model
Mingcai Zhang, Michigan State University; Lihong Yang, Australian Curriculum, Assessment and Reporting Authority

To evaluate the impact of construct-shift in vertical scaling, the unidimensional and multidimensional IRT models are compared in developing a single vertical scale across grades. Simulation conditions include construct correlation, average achievement growth, and anchor item design. The reference composite based on MIRT approach performs better in improving score comparability.

Measurement Invariance for Common Core State Standards Aligned Computerized Adaptive Tests
Xueming Li, Northwest Evaluation Association

This study examines measurement invariance of CCSS-aligned computerized adaptive tests. Longitudinal data from various states are analyzed. The results provide important evidence of measurement invariance for CCSS-aligned tests and empirical evidence for measuring student growth.

Monitoring Scale Stability of a CAT program Using Quality Control Charts
Wei He, NWEA; Jiahui Zhang, Michigan Stage University

This study explored the use of statistical quality control methods including Shewhart and CUSUM charts in monitoring scale stability in a large-scale vertically-scaled CAT program. Impacts of different sample size on these quality control charts were also examined.

Multiple Group IRT Vertical Scaling in Measuring Student Growth
Yanming Jiang, Educational Testing Service

This study is focused on vertical scaling for test populations that consist of heterogeneous groups within each test level/grade. Four multiple group vertical scaling methods are proposed and will be evaluated in terms of the accuracy of established vertical scales and cohort growth trajectories measured by the vertical scales.

Sunday, April 30, 2017
10:35 AM–12:05 PM, Salon K, Meeting Room Level, Paper Session, J7

Testlet-based Tests

A Random Item Mixture IRT Testlet Model for Speededness in CR Items

Meereem Kim, University of Georgia; Allan Cohen, University of Georgia

Time limits on test have been found to result in speededness effects, negatively affecting item parameter estimates in fixed item models. In this study, a random item mixture IRT testlet model is investigated for detection of speededness effects in constructed response items. Results are compared with a fixed item model.

Applying Rasch Testlet Models to CAT with Varied Testlet Characteristics

Seohong Pak, The University of Iowa; Hong Qian, NCSBN

This simulation study was designed to investigate the impacts of different testlet characteristics, such as a testlet size, a testlet composition, and a testlet random effect, on estimating individual's ability under CAT comprised only of testlet-based items.

Comparing DCM and Bayes Nets Approaches for Modeling Testlet-based Data

Meghan Sullivan, University of Kansas; Jonathan Templin, University of Kansas

This research aims to 1) compare BayesNets and Diagnostic Classification Models and 2) replicate and extend previous research involving testlet-structures modeled with these methods. Misspecification of model measurement and structural components and complex testlet-item structures will also be investigated.

Optimizing the Multi-phase Sampling in Selecting Testlets with Constraints on Reused Blocks

Jiahe Qian, Educational Testing Service; Shuhong Li, Educational Testing Service; Lixiong Gu, Educational Testing Service

In sampling testlets consists of reused item-blocks, for security reasons, strict constraints are often imposed to avoid overusing the same blocks. Multi-phase sampling is proposed to select as large a sample of testlets as possible and optimized algorithms are applied to improving the psychometric properties of the testlets sampled.

Sunday, April 30, 2017
10:35 AM–12:05 PM, Salon L, Meeting Room Level, Paper Session, J8

Subscore Reliability

Session Chair: Yu Fang, ACT, Inc.

Can Subtest Equating Borrow Information from the Full Test?

Yu Fang, ACT, Inc.; Yang Lu, ACT, Inc.; Yong He, ACT, Inc.

This study compares three approaches to subscore equating under the random equivalent groups design: the equipercentile method using subscores, the chained equipercentile method using equated total scores as the anchor, and the IRT equating method using scaled item parameter estimates on full tests.

Evaluation of Augmented Subscore Equating or Scaling Methods

Yuming Liu, Educational Testing Service

Real and simulated data were used to evaluate the psychometric features of augmented subscore equating or scaling in terms of value-added and difference between augmented and number-correct subscore equivalents. The impacts of sample size, number of linking items, and correlation among subscores on augmented subscore equating/scaling were also examined.

Improving Subscale Score Reliability in Large-scale Assessments

HeaWon Jun, Georgia Institute of Technology; Susan Embretson, Georgia Institute of Technology

The current study compares the reliability of several subscale scoring methods over varying data conditions. Both classical test theory and item response theory methods for subscale scoring are compared across simulated data conditions, varying in subscale length, subscale inter-correlations, item inter-correlations, and level of item difficulty.

Investigating Subscore Accuracy under Various CAT Designs

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

Considering the increasing demand on reporting subscores for diagnostic and remedial purposes, decent subdomain ability estimations should be required. This simulation study investigates and compares the performance of different CAT designs in minimizing the standard error of subdomain ability estimates and thus achieving desired subscore accuracy under various conditions.

Reliability of Mixed-Format Composite Scores Involving Raters: A Multivariate Generalizability Theory Approach

Stella Kim, The University of Iowa; Won-Chan Lee, The University of Iowa

The primary purpose of this study is to estimate mixed-format composite score reliability using multivariate generalizability theory involving a rater facet. Several D studies are carried out by varying the number of items and raters. Various mixed-format data are used with different test characteristics such as section weights and dimensionality.

Sunday, April 30, 2017
10:35 AM–12:05 PM, Salon M, Meeting Room Level, Paper Session, J9

Applications of Structural Equation Models

Chair: Nicole Brocato, Wake Forest University

Assessing the precursors and attainment of wellbeing in higher education

Nicole Brocato, Wake Forest University; Eranda Jayawickreme, Wake Forest University; John Pryor, Pryor Education Insights

We propose that measures of wellbeing should assess not only the attainment of wellbeing, but also whether respondents have the necessary precursors for wellbeing. Using structural equation modeling, we present results from an ongoing effort to develop a measure that assesses wellbeing, its precursors, and its correlates.

Exploring Reliability of Each Area for a Multistage Adaptive Test

Xinhui Xiong, American Institute of Certified Public Accountants; Seohyun Kim, The University of Georgia

This study compares the reliability estimate of a licensure exam with coefficient alpha and the nonlinear SEM reliability coefficient. Latent classes of a dataset from the exam will be examined, and the reliability estimates will be obtained and compared for the latent classes and manifested groups.

Factor Structure of Mathematic Attitudes for Five Asian Countries on TIMSS-2011

Peiyan Liu, University of Denverr

Using eighth grade students' data from TIMSS-2011, the factor structure, measurement invariance, and validity will be investigated for the four scales of mathematic attitudes: confidence, engagement, like mathematics, and value mathematics. Data from five Asian countries will be included in this analysis: Hong Kong, Chinese Taipei, Japan, Korea, and Singapore.

Measurement Invariance of Non-cognitive Assessment in a Workplace Setting

JiYoon Kim, Talent Assessment Institute (TAI); Hanwook Yoo, Educational Testing Service; Youngseok Lee, Talent Assessment Institute (TAI); Myungkyu Lee, Talent Assessment Institute (TAI); Heejae Yoo, Talent Assessment Institute (TAI); Davaasuren Tumendemberel, Talent Assessment Institute (TAI)

This study evaluates the invariance in the factor structure of a non-cognitive assessment in a workplace setting. The subgroups are defined by employment status and nationality. Findings from this study provides how to reduce the impact of cultural differences when the test is adapted from different linguistic and cultural context.

Using Structural Equation Modeling to Examine Position Effects in 2PL IRT Model

Wei Tang, University of Alberta; Qi Guo, University of Alberta; Okan Bulut, University of Alberta

Item position effects are often assessed using Hierarchical Generalized Linear Model, which is equivalent to the Rasch model. This study proposed a Structural Equation Modeling approach to test item position effects in two-parameter IRT model. A Monte Carlo simulation study examines position effects in item difficulty, item discrimination, and both.

Sunday, April 30, 2017
10:35 AM–12:05 PM, Salon J, Meeting Room Level, Electronic Board Session, Paper Session, J10

Electronic Board #1

Quantifying the Item Order Effect on Item Difficulty in Large-Scale Testing

Kuan Xing, University of Illinois at Chicago; Kirk Becker, Pearson VUE

While many studies exploring the effect of item order, testing professionals remain concerned about changing or randomizing item order. For this study item difficulty is computed and compared based on the relative position where the item was administered. Strong evidence shows no significant effect of item order on item difficulty.

Electronic Board #2

Differential Item Functioning Effect Size Use for Validity Information

Brian French, Washington State University; Maria Dolores Hidalgo Montesinos, University of Murcia; Holmes Finch, Ball State University; Maria Hernandez Finch, Ball State University

The majority of assessments contain differential item functioning (DIF). A practitioner must use an assessment containing as little DIF as possible. Several effect size indices are evaluated to capture aggregate DIF in assessments with a focus on random effects (RE) models, as RE effect sizes may capture DIF variability best.

Electronic Board #3

Automated Modeling of Item Difficulty on Admissions Tests

Kirk Becker, Pearson; Rebecca Mathew, Illinois Mathematics and Science Academy; James Masters, Pearson

This investigation identified and extracted semantic features from test content on an admissions test, and used those features to explore item difficulty. Both linear multiple regression and mixture models were used to model item difficulty. Results were compared with SME judgements of item difficulty.

Electronic Board #4

Semantic Similarity Item Analysis for Short-Answer Questions

Ryoungsun Park, Wayne State University; Hyewon Chung, Chungnam National University; Jiseon Kim, University of Washington; Barbara Dodd, University of Texas

This study presents an item analysis for short-answer questions using natural language processing. A single metric of semantic variability, which indicates the quality of an item, was derived using the word-answer frequency table and Kaiser-Meyer-Olkin (KMO) index. The results of the item analysis of students' actual answers are demonstrated.

Electronic Board #5

Can Task Models be used to Improve Equating?

Jaime Malatesta, University of Iowa; Huan Liu, University of Iowa

This study evaluates the interchangeability of free response items written from task models by using a variety of procedures originally designed to detect unstable anchor items. The best performing free response items are then treated as anchor items and results are compared against MC-only anchor sets for equating mixed-format tests.

Electronic Board #6

Investigating "Classification Equity" Property of Equating

Shalini Kapoor, ACT; Yi-Fang Wu, ACT, Inc

An equity property of equating desirable for equated tests with cut scores is that examinees at a particular ability level, given a cut score, should have the same probability of passing/failing on the two test forms. This equity property referred to as "classification equity" is evaluated in this paper.

Electronic Board #7

Construct Comparability between Paper- and Digital-Based Assessments in NAEP Grade 8 Mathematics

Young Yee Kim, American Institutes for Research; Soo Lee, American Institutes for Research; Jiao Yu, American Institutes for ResearchNAEP is in a transition to digitally-based assessments (DBAs) from a paper-based assessments (PBAs). The possible introduction of digital familiarity as a nuisance construct raises concerns about the comparability of the two modes of assessment. This study examines construct comparability between PBAs and the DBA for grade 8 mathematics.

Electronic Board #8

Gathering and Evaluating Validity Evidence: The Generalized Assessment Alignment Tool

Gregory Cizek, University of North Carolina at Chapel Hill; Audra Kosh, University of North Carolina at Chapel Hill; Emily Toutkoushian, University of North Carolina at Chapel Hill

We present a new method for aligning tests to content standards: the Generalized Assessment Alignment Tool (GAAT), applicable to diverse test purposes, educational and credentialing contexts, administration modes, and item formats. The GAAT produces four indices addressing three aspects of alignment: Construct Comprehensiveness, Content Concentration, and Cognitive Complexity.

Electronic Board #9

Monitoring the equating transformation to ensure quality over time

Marie Wiberg, Umeå University

The aim was to examine the equating transformations over a large number of administrations of a college admission test in order to ensure quality over time. Different equating methods, test groups and different linking plans are examined. Recommendations for performing equating consecutively over several administrations in the future are given.

Electronic Board #10

Critical thinking and self-assessment using formative assessment with grade 4 students

Beverly FitzPatrick FitzPatrick, Memorial University; Henry Schulz, Memorial University

Young students should be taught to think critically. Self-assessment is important to students learning to think independently and critically. We worked with grade 4 students to improve their critical thinking and self-assessment using formative assessment. Through quantitative and qualitative analysis of four formative assessments, we determined how students learned.

Electronic Board #11

Comparing Propensity Score Matching and Weighting for Random Groups EquatingAccuracy

YoungKoung Kim, The College Board; Tim Moses, The College Board; Judit Antal, The College Board

Propensity score matching and weighting procedures have been proposed to approximate randomly equivalence from systematically different equating groups. The goal of the present study is to compare the effectiveness of both procedures for obtaining randomly equivalent groups and improving the accuracy of random groups equating.

Electronic Board #12

Exploratory Text Analysis of DIF Items with Common Reference Groups

Natalie Jorion, Pearson VUE; Kirk Becker, Pearson VUE; Joseph Betts, Pearson VUE; Ada Woo, National Council of State Boards of Nursing

Given the complexity of identifying potential reasons for flagged DIF items, this research explores a complementary method for evaluating DIF using natural language processing. This investigation would calculate the degree of textual similarity between items with common reference groups compared to DIF and non-DIF items using a data mining approach.

Electronic Board #13

Incorporating A-stratified Strategy in Variable-length Computerized Classification Testing

Xiao Li, University of Illinois at Urbana-Champaign; Huahua Chang, University of Illinois at Urbana-Champaign

An adaptation of the a-stratified method (Chang and Ying, 1999) is proposed for variable length computerized classification testing. Our pilot study shows that the method not only yields equivalent classification consistency and accuracy, but also more balanced exposure control with respect to some traditional methods.

Electronic Board #14

Using the odds ratio to detect differential item functioning

Kuan-Yu Jin, Education University of Hong Kong; Wen-Chung Wang, Education University of Hong Kong; Hui-Fang Chen, City University of Hong Kong

We develop a simple but efficient method for DIF assessment using the odds ratio of two groups of examinees by two groups' responses to a studied item. Simulation results demonstrated the superiority of the proposed approach to the Mantel-Haenszel and logistic regression methods, especially when there were missing data.

Electronic Board #15

Effects of Item Parameter Drift in a Large-scale Assessment on Educational Research

HyeSun Lee, California State University Channel Islands; Kurt Geisinger, University of Nebraska-Lincoln

This research examined the effects of Item Parameter Drift occurring in a short scale from a large-scale assessment on parameter estimates in multilevel models where scale scores were employed as time-varying predictors to account for educational achievement. Practical implications of the findings were discussed to improve accuracy in educational research.

Electronic Board #16

Predictors of Low Agreement Between Automated Speech Recognition and Human Scores

Joseph Nese, University of Oregon; Josh Kahn, University of Oregon; Akihito Kamata, Southern Methodist University

The purpose of this study is to investigate potential student and passage predictors of low agreement between an automated speech recognition (ASR) engine and human scores of words read correctly in student oral reading fluency passages. We fit a multi-level, cross-classified IRT model to model a latent estimate of agreement.

Electronic Board #17

Predicting 8th Grade Math Proficiency with 1st and 3rd Grade Math Performance

Se-Kang Kim, Fordham University; Jessica Lutz, Fordham University

The study predicted 8th grade math proficiency given 1st and 3rd grade math performance. The results showed the 1st grade math achievement was strongly related with 8th grade math proficiency, implying that school personnel may put more efforts on earlier grade math education for the better outcome for advanced grades.

Electronic Board #18

An Antedependence Model for Longitudinal Item Response Data

XUE ZHANG, NORTHEAST NORMAL UNIVERSITY; CHUN WANG, University of Minnesota at Twin Cities; Jian Tao, Northeast Normal University

A first-order antedependence growth model is proposed for longitudinal data with binary indicators. It relaxes the stationarity assumptions of autoregressive models because residuals of traits over time are often non-stationary. The performances of antedependence models and autoregressive models are compared in a simulation study.

Electronic Board #19

Modeling the Dynamic Nature of Student Learning: A Systems Approach

Pamela Paek, ACT; Britte Cheng, SRI; Paul Nichols, ACT; Geneva Haertel, SRI

This project explores and illustrates the value of systems models to explore the complex, dynamic nature of educational policies and practices, including key players, factors, and interactions and relationships that are currently not modeled or measured in educational research.

Electronic Board #20

What Is Process Data?

Andreas Oranje, Educational Testing Service

Process data seems to be the future of assessment if you count how often those and similar words surface. Yet, there seems very little agreement about what it exactly is. This paper attempts to build a common understanding about uses, design, types, analysis, and applications of process data.

Electronic Board #21

Test design considerations for reporting subscores in large-scale educational survey assessment

Nuo Xi, Educational Testing Service; Yue Jia, Educational Testing Service; Xueli Xu, Educational Testing Service; Longjuan Liang, Educational Testing Service

The research objective is to investigate the impact of varying content area subscale length per examinee (especially when it is short) for its prospective use in large-scale educational survey assessments. Simulation studies and a sensitivity study using real data will be presented to evaluate the effect on estimating group-level statistics

Electronic Board #22

Using Probability Estimates to Reduce Pretest Sample Sizes

Jerome Clauser, American Board of Internal Medicine; Gerald Arnold, American Board of Internal Medicine

To ensure item quality, testing organizations routinely pretest items before they appear on a live form. Although valuable, this practice comes with a significant cost in testing time. This manuscript uses probability estimates to classify items. Results indicate that item quality can be maintained with significantly fewer exposures per item.

Electronic Board #23

Estimation of SGP measurement errors using test characteristic functions

Jinah Choi, The University of Iowa; Robert Ankenmann, The University of Iowa; Won-Chan Lee, The University of Iowa

This study presents an analytic approach for estimating conditional standard errors of measurement and reliability for Student Growth Percentiles by using test characteristic functions in unidimensional IRT. The proposed method is applied to data from a statewide assessment, and results are compared to various other alternative procedures and scoring methods.

Sunday, April 30, 2017
12:00 PM–2:00 PM, Conference Room 9, Meeting Room Level

NCME Past President's Luncheon.

By Invitation Only.

Sunday, April 30, 2017
12:25 PM–1:55 PM, Salon A, Meeting Room Level, Coordinated Session, K1

Multi-Method Validation Using an Integrated Evidence-Centered Design Process

Session Chair: Drew Gitomer, Rutgers University Graduate School of Education

Traditional and even newer approaches to validation, such as evidence-centered design (ECD), have focused on the validity argument and relevant evidence addressing the inferences and use of scores derived from specific assessments. This session extends ECD research by using an integrated ECD process to build a broad set of measures within the same construct domain and to create and evaluate a coordinated validity argument that examines multiple measures within that domain. We explore the fundamental question that underlies the construct of Content Knowledge for Teaching (CKT)—what does a teacher know about teaching a particular content area, and how does that knowledge relate to instructional quality and student learning in that content area? Within a unified evidence model and a single competency model, we can examine both the possession and enactment of CKT. Linking different sources of evidence from demonstrations of teachers' knowledge (CKT-E assessment) and their enactment of teaching practices (observation and artifact assessments) within a single framework guides the design of these different measurement instruments and supports enriched theoretical interpretations. We go beyond correlational studies and examine how teachers with differing knowledge profiles enact instruction and how, together, this influences student learning.

A Validity Approach Integrated ECD Design: Framework and Methodology
Drew Gitomer, Rutgers University Graduate School of Education; Courtney Bell, Educational Testing Service (ETS); Jim Minstrell, Facet Innovations; Ruth Anderson, Facet Innovations

Design and Development of a CKT-E Assessment
Geoffrey Phelps, Educational Testing Service (ETS); Drew Gitomer, Rutgers University Graduate School of Education; Charles Iaconangelo, Rutgers University Graduate School of Education

Cognitive Diagnostic Modeling and Validation Using Known Groups
Charles Iaconangelo, Rutgers University Graduate School of Education; Geoffrey Phelps, Educational Testing Service (ETS); Drew Gitomer, Rutgers University Graduate School of Education

Classroom Observations as a Measure of the Enactment of CKT
Courtney Bell, Educational Testing Service (ETS); Robert Zisk, Rutgers University Graduate School of Education; Drew Gitomer, Rutgers University Graduate School of Education

Classroom Artifacts as a Measure of Enactment of CKT
Robert Zisk, Rutgers University Graduate School of Education; Drew Gitomer, Rutgers University Graduate School of Education; Courtney Bell, Educational Testing Service (ETS)

Sunday, April 30, 2017
12:25 PM–1:55 PM, Conference Room 1&2, Meeting Room Level, Coordinated Session, K2

Bayesian Sequential Methods for Adaptive Testing

Session Organizer: Wim van der Linden, Pacific Metrics Corporation
Session Chair: Michelle Barrett, Pacific Metrics Corporation
Session Discussant: Deborah Harris, ACT, Inc.

In adaptive testing programs, the item parameters are typically fixed at point estimates and subsequently treated as known. In this symposium, Bayesian methods are presented to account for the actual uncertainty in the item parameters, to update their posterior distributions during operational testing, and deal with the scoring of prematurely ended tests. The first two methods are based on the optimized MCMC algorithm for adaptive testing presented in van der Linden and Ren (2015, 2016). Because of immediate mixing of the Markov chain and extremely simple posterior calculations, the algorithm selects items from real-world-size item pools in milliseconds. The last method uses the idea of sampling from posterior predictive distributions ("multiple imputations") to evaluate several of the current penalty-based scoring methods for incomplete tests.

Bayesian Adaptive Testing with Polytomous Items
Hao Ren, Pacific Metrics Corporation; Seung Choi, Pacific Metrics Corporation; Wim van der Linden, Pacific Metrics Corporation

Bayesian Ability Estimation in Operational Adaptive Testing
David King, Pacific Metrics Corporation; Wim van der Linden, Pacific Metrics Corporation

Evaluation of Penalty-Based Scoring Methods for Incomplete Adaptive Tests
Qi Diao, Pacific Metrics Corporation; Wim van der Linden, Pacific Metrics Corporation

Shadow-Test Approach to Joint Adaptive Testing and Item Calibration
Wim van der Linden, Pacific Metrics Corporation; Bingnan Jiang, Pacific Metrics Corporation

Sunday, April 30, 2017
12:25 PM–1:55 PM, Conference Room 3&4, Meeting Room Level, Coordinated Session, K3

New measures of digital access, familiarity, efficacy, and engagement for large-scale assessments

Session Organizer: Sami Kitmitto, American Institutes for Research
Session Chair: George Bohrnstedt, American Institutes for Research
Session Discussant: Lauren Harrell, National Center for Education Statistics, Assessments Division

As assessments move from paper-and-pencil administrations to digitally-based administrations an important issues is the degree to which all children are ready for this change. This session will focus on the development of measures of access, familiarity, self-efficacy, and engagement with digital technology in two major assessment programs – the National Assessment of Educational Progress (NAEP) and the Programme for International Assessment (PISA). Two papers will discuss new approaches and findings in the development of items to measure these constructs from a special NAEP study conducted in 2015. The first paper will discuss new, innovative survey methods for improving the measurement of familiarity with using vignettes to anchor self-reporting scales and foil concepts to address issues of overclaiming. The second paper will use the larger set of items included in the special study to discuss the construction of indices to measure access, familiarity, and self-efficacy and how the indices are distributed across subpopulations (gender, race/ethnicity, eligibility for the National School Lunch Program). The third and final paper will use PISA data to examine the motivational structure of information and communication technology engagement and whether log-data in computer-based assessments can explain additional variance in the prediction of relevant criteria.

Measuring familiarity with digital technology using innovative item formats
Jonas Bertling, Educational Testing Service; Debby Almonte, Educational Testing Service

Developing New Indices to Measure Computer Access, Familiarity, and Self-efficacy
Sami Kitmitto, American Institutes for Research; Bitnara Park, American Institutes for Research

Relationship between information and communication technology (ICT) usage and ICT engagement
Olga Kunina-Habenicht, German Institute for International Educational Research (DIPF), Centre for International Student Assessment (ZIB); Frank Goldhammer, German Institute for International Educational Research (DIPF), Centre for International Student Assessment (ZIB)

Sunday, April 30, 2017
12:25 PM–1:55 PM, Salon B, Meeting Room Level, Coordinated Session, K4

Current Developments in Selective Admission to Higher Education in Europe

Session Chair: Susan Niessen, University of Groningen
Session Discussant: Brent Bridgeman, Educational Testing Service (ETS)

Globally, there is an increasing interest in broadening admission criteria for higher education. The aims of using criteria broader than general educational achievement is to improve predictive validity, to predict alternative outcomes such as later job performance, and to improve accessibility for minority applicants. This session contains four studies that show European efforts to meet these objectives. Wendy de Leng discusses the development of a situational judgement test for medical school applicants measuring integrity-related traits. Construct validity, a major concern with SJTs, was studied in two samples. Karen Stegers-Jager studies the relationship between ethnic and social background and acceptance rates into medical school. Non-academic criteria did not result in significantly different acceptance chances, but academic criteria did for some groups. Susan Niessen studies the impact of self-presentation on the scores, validity and hiring decisions based on non-cognitive self-report scales administered in an admission context. Scores obtained in an admission context were significantly different from scores obtained in a low-stakes context and predictive and incremental validity were attenuated. Rob Meijer discusses the predictive validity, and applicant reactions for a work sample approach in selective admission, using trial-studying tests. Implications for admission practices and future directions based on these studies are discussed.

Hybrid Development of an Integrity-Based Situational Judgment Test for Medical School Selection
Wendy de Leng, Institute of Medical Education Research Rotterdam (iMERR), Erasmus Medical Center; Karen Stegers-Jager, Institute of Medical Education Research Rotterdam (iMERR), Erasmus Medical Center; Marise Born, Erasmus University Rotterdam; Axel Themmen, Erasmus Medical Center

Applicants' Ethnic and Social Background and Performance on Different Selection Criteria
Karen Stegers-Jager, Institute of Medical Education Research Rotterdam (iMERR), Erasmus Medical Center

Using Self-Report Questionnaires in High-Stakes Contexts
Susan Niessen, University of Groningen; Rob Meijer, University of Groningen; Jorge Tendeiro, University of Groningen

A Work Sample Approach to Selection and Matching in Higher Education
Rob Meijer, University of Groningen; Susan Niessen, University of Groningen

Sunday, April 30, 2017
12:25 PM–1:55 PM, Salon C, Meeting Room Level, Coordinated Session, K5

The Impact of Background Knowledge in Reading for Understanding

Session Chair: Jonathan Weeks, Educational Testing Service
Session Discussant: James Pellegrino, University of Illinois at Chicago

In recent years the construct of reading comprehension (RC) has expanded to include content area and disciplinary reading which suggests that students should read differently depending upon the discipline (e.g., history vs. science; Goldman, 2012). This approach to reading conflicts with many existing assessment frameworks where content area texts are purposely avoided in an attempt to reduce the influence of background knowledge. Increasing the content and topical specificity of the texts in a reading assessment has the potential to increase the influence of background knowledge (BK) on RC scores (Shapiro, 2004). However, it is unclear if comprehension scores reflect underlying reading ability or simply students' BK levels. The papers in this session are intended to address several foundational questions related to the impact of BK on comprehension. 1) Should each RC test have its own topically related BK measure? 2) When should students' BK be measured (before or after the RC items) such that BK and RC scores are minimally impacted? 3) Is BK psychometrically distinct from RC, and is there evidence of disciplinary literacy? 4) Is there an optimal level of BK required to comprehend text? 5) Does BK have an impact on comprehension for beginning readers?

The problem of background knowledge in measuring reading comprehension
Tenaha O'Reilly, Educational Testing Service

Does background knowledge prime performance on cognitive test items?
Jonathan Steinberg, Educational Testing Service; Jennifer Minsky, Educational Testing Service

Distinguishing between background and cognitive knowledge: Examining domain-specific factors in reading performance
Jonathan Weeks, Educational Testing Service

A regression-based approach to identifying a knowledge threshold in reading comprehension
Zuowei Wang, Educational Testing Service

The effect of background knowledge on comprehension for K-3 students
John Sabatini, Educational Testing Service; Laura Halderman, Educational Testing Service

Sunday, April 30, 2017
12:25 PM–1:55 PM, Salon D, Meeting Room Level, Paper Session, K6

Teacher Evaluation

Measurement Error and Bias in Value-added (VAM) Scores

Michael Kane, Educational Testing Service

Measurement errors in prior scores add bias to residual gain scores for students, and thereby, to VAM scores for teachers and schools. This bias can be substantial, but it can be minimized by an appropriate statistical correction. The correction requires accurate estimates of generalizability coefficients reflecting all sources of error.

Non-parametric Estimation of Student Growth Percentile

Ruitao Liu, ACT

The B-spline based quantile regression method for calculating Student Growth Percentile can produce biased estimations when model misspecification happens. We propose several non-parametric approaches to get more reliable estimations. A simulation study was performed to compare the proposed methods and the standard approach through their empirical finite sample properties.

Using Large-Scale Student Ratings of Instruction for Improving Evaluation of College Teaching

Stephen Benton, The IDEA Center; Dan Li, The IDEA Center; Jason Barr, The IDEA Center

Using Bayesian Model Averaging, we analyzed course-level student ratings of instruction (SRI) data collected in undergraduate and graduate classes (N = 6,405) from 27 institutions, representing all regions of the continental U.S. Distinctive patterns of teaching methods emerged as significant predictors for student progress on each relevant learning objective.

Value-Added Modeling without SAT/ACT Scores

Jonathan Lehrfeld, Council for Aid to Education; Eric Muller, Council for Aid to Education; Doris Zahner, Council for Aid to Education

We discuss the difficulty of using a value-added model that relies on SAT scores to correct for differing entering academic ability. We present a model that uses covariates without missing data, and address the questions of how the new model affects value-added scores and whether the new estimates are reliable.

Sunday, April 30, 2017
12:25 PM–1:55 PM, Salon K, Meeting Room Level, Paper Session, K7

Linking & Equating with Small Sample Sizes

Session Chair: Darius Taylor

A Bayesian Historical Data Borrowing Approach for Small-Sample Linear Equating

Jianshen Chen, Educational Testing Service; Wei Wang, Educational Testing Service; Yi Cao, Educational Testing Service; James Morgan, Educational Testing Service

Making full and appropriate use of available data is critical in small-sample equating. A Bayesian approach is proposed to dynamically incorporate historical information for obtaining the posterior mean and SD of a new form in linear equating. Results showed that this approach can increase equating accuracy and reduce equating error.

Comparing Circle Arc and Nominal Weights Mean Equating with Small Samples

Lisa Keller, University of Massachusetts Amherst; Rob Keller, Measured Progress; Michael Nering, Measured Progress; Andrea Hebert, Quaternion Group

While some research recommends no equating when sample sizes are small, the results of this study suggests for samples with at least 75 examinees, equating with either method was better than no equating. Circle arc produced more bias estimates than nominal weights, which generally produced better results.

Equating under Small Samples: A Comparison of Six Methods

Fen Fan, National Commission on Certification of Physician Assistants; Drew Dallas, National Commission on Certification of Physician Assistants; Joshua Goodman, National Commission on Certification of Physician Assistants

This study examines the performance of several new small-sample equating methods (circle-arc, nominal weights mean, & general linear methods) under NEAT design in terms of equating accuracy and decision consistency/accuracy using both simulated and real data. Results showed all three methods are promising at sample sizes as low as 25.

Expected Classification Error under Small Sample Equating with Mixed-Format Tests

Ja Young Kim, TEPS Center Seoul National University

Few studies investigated equating with small samples using mixed-format tests. The purpose of this study is to examine the impact of small sample and equating method on the misclassification of examinees based on where the passing scores are located, taking into account factors related to using the mixed-format tests.

IRT equating versus classical equating for small sample test forms

Andrew Dwyer, The American Board of Pediatrics; Robert Furter, The American Board of Pediatrics

This study uses both a simulation component and a resampling component to compare the efficacy of IRT equating (anchored Rasch calibration) and classical equating (circle arc, nominal weights mean) methods for maintaining equivalent classification decisions across small sample test forms (N = 10, 25, 50, 100).

Sunday, April 30, 2017
12:25 PM–1:55 PM, Salon L, Meeting Room Level, Paper Session, K8

Technical Advances in IRT

A Finite Mixture IRT Model for Continuous Measurement Outcomes

Cengiz Zopluoglu, University of Miami

A mixture extension of Samejima's Continuous Response Model for continuous measurement outcomes and its estimation through a heuristic approach based on limited-information factor analysis is introduced. The effectiveness of this estimation approach under real data analytic conditions was examined through a Monte Carlo simulation study and an empirical data demonstration.

A method for identifying linear item position effects on 3PL-IRT item parameters

Min Sung Kim, Buros Center for Testing; Chansuk Kang, University of Nebraska-Lincoln; Anthony Albano, University of Nebraska-Lincoln

This study advances methods for identifying item position effects, which currently only detect changes in item difficulties, by including the detection of item discrimination and guessing parameter shifts. The method was applied to PISA 2009 reading data to examine the effect of moving items to later positions on the test.

An Improved Version of Method A Based on Saddlepoint Approximation

Yinhong He, Beijing Normal University; Ping Chen, Beijing Normal University; Yong Li, Beijing Normal University; Shumei Zhang, Beijing Normal University

By combining saddlepoint approximation (Daniels, 1954) with Method A (Stocking, 1988), this study proposed a new online calibration method in CAT, named saddlepoint approximation-Method A (SA-Method A), with the intention to weaken the uncertainty of ability estimates in Method A and further improve its calibration precision.

An IRT Model with Ability-Difficulty-Discrepancy-Based Guessing Coefficient

Luping Nlu, The University of Texas at Austin

This study proposes a 3PL-ADGC model to account for the impact of guessing, represented by the product of a guessing coefficient and the discrepancy between person's ability and item difficulty. The new model is found to fit better than other IRT models tested in a real dataset.

Threshold-Autoregressive Item Response Theory (TAR-IRT)

Xiaodan Tang, University of Illinois at Chicago; George Karabatsos, University of Illinois at Chicago; Haiqin Chen, American Dental Association

Empirical violations of the local independence (LI) assumption bias the parameter estimates of standard IRT models. We propose a TAR-IRT model, characterized by a less strict form of LI that additionally accounts for time dependence parameters. We illustrate our modeling approach through the analysis of a large item-response data set.

Sunday, April 30, 2017
12:25 PM–1:55 PM, Salon M, Meeting Room Level, Paper Session, K9

Reliability Issues

Session Chair: Dvir Kleper, National Institute for Testing and Evaluation

A Comparative Study on Methods for Estimating Scale Score Reliability

Chen Li, University of Maryland, College Park; Tim Moses, The College Board; YoungKoung Kim, The College Board; Weiwei Cui, The College Board; Amy Hendrickson, The College Board

This study compares scale score conditional standard error of measurement (CSEM) and reliability estimates yielded from the compound binomial method (Kolen et al., 1992), the Item Response Theory (IRT) method (Kolen et al., 1996) and the generalized alpha method (Almehrizi, 2013) using a real dataset from a large scale assessment.

Adaptive Monitoring of Rater Effects --Two New Essay Selection Methods

Zhuoran Wang, University of Minnesota, Twin Cities; Chun Wang, University of Minnesota, Twin Cities

Two new essay selection methods were proposed to assess rater leniency effect and centrality effect. The Leniency method resulted in accurate rater leniency estimate. The Centrality method had precise rater centrality estimate, especially in long tests. The Centrality method resulted in the most skewed essay bank usage.

Comparing the Accuracy of Student Growth Measures

Katherine Castellano, Educational Testing Service; Daniel McCaffrey, Educational Testing Service

With new tests aligned to the Common Core and less emphasis on universal proficiency by ESSA, states are reconsidering their choice of growth measure. This paper presents relative accuracy findings for common student growth measures to help states determine which measure will provide the most meaningful information for their needs.

Retaking the Psychometric Entrance Test

Dvir Kleper, National Institute for Testing and Evaluation; Elliot Turvall, National Institute for Testing and Evaluation

Using a two level hierarchical model (also known as a growth model) we model a longitudinal data in order to characterize the pattern of retests among examinees who take the Psychometric Entrance Test (PET) a number of times in relation to their demographic background variables.

Testing classification using the method of boosting

Yating Zheng, University of Maryland at College Park; Kecheng Xu, University of Texas at Austin

This study explores boosting, a popular classification method in machine learning, in testing classification. Comparison is made between the classification accuracy using the Rasch model and that using the method of boosting. Results indicate that boosting provides better classification accuracy than the Rasch model when the test volume is low.

Sunday, April 30, 2017
12:25 PM–1:55 PM, Salon J, Meeting Room Level, Electronic Board Session, Paper Session, K10

Electronic Board #1

Learning Promotion and Mastery Detection of Multiple Attributes

Sangbeak Ye, University of Illinois - Urbana Champaign; Jeff Douglas, University of Illinois - Urbana Champaign

Bayesian item selection methods and sequential mastery detection methods are used conjointly to construct an e-learning assessment that aims for mastery of multiple attributes. Under cognitive diagnosis models, the introduced methods are meant to provide a self-guided assessment with minimal number of items at a fixed false detection rate.

Electronic Board #2

Application of Generalizability Theory to Classroom Assessments in a School Accountability Context

Susan Lyons, National Center for the Improvement of Educational Assessment; Carla Evans, University of New Hampshire

This study examines the reliability of generalization from a collection of classroom assessments intended to measure student achievement to the universe of all possible assessments. It also determines an efficient number of classroom assessments necessary to ensure high reliability of estimates of student achievement made in a school accountability context.

Electronic Board #3

Constructing a Decision Making Assessment for College Admission: a Psychometric Perspective

Brad Wu, Pearson VUE

The aim of this study is to create a decision making test which will provide a broad sampling of the cognitive processes associated with problem solving, dealing with uncertainty, and managing risks to enhance the admission process. Psychometric properties such as reliability, item types, timing and scaling are examined.

Electronic Board #4

Fairness in Testing: Evaluating the Alignment of Standards, Guidelines, and Practices

Jessica Jonson, Buros Center for Testing - University of Nebraska-Lincoln; Betty-Jean Usher-Tate, Buros Center for Testing - University of Nebraska-Lincoln

This presentation will share results from a content analysis aligning standards and guidelines for test fairness with actual practices as documented in test manuals and published reviews of educational tests. Potential gaps between standards, guidelines, and practices as well as the identification of notable practices will be discussed.

Electronic Board #6

Test designs and modeling under the nominal responses diagnostic model framework

Jinsong Chen, SYSU

Under the nominal response diagnostic model, there are different ways to design situational tests. The saturated model subsumes different reduced forms that can help to inform if the test is designed as expected. Two simulation studies were adopted to demonstrate the effectiveness of the models and designs.

Electronic Board #7

Independent Replicability of Test Developers' Item Content Classifications: A Meta-Analysis

Anne Traynor, Purdue University

The extent to which a test developer's item content classifications can be confirmed by independent subject-matter experts is an important aspect of test content alignment evidence. We use meta-analysis to identify alignment procedure features that appear to influence agreement between independent experts' and test developers' item-objective matches.

Electronic Board #8

Development and Validation of a Social Emotional Learning Screening Assessment for Children

Stephen Elliott, Arizona State University; Michael Davies, Griffith University; Jennifer Frey, George Washington University

Social emotional learning is a critical aspect of schooling. This study examined the validation of a new universal screening measure of SEL that is based on the CASEL Five model. The results indicated the SELA is an efficient, reliable teacher completed measure that discriminates students at-risk socially and academically.

Electronic Board #9

A Predictive Validity Study: Do State Assessment Scores Predict ACT Scores?

Jie Chen, University of Kansas; Wenhao Wang, University of Kansas

This study explores relationships between a state assessment scores and ACT scores to see if students' performance on the state assessment could successfully predict their achievement on ACT as well as their college success. It further examines these relationships across gender and ethnicity groups to see if they change.

Electronic Board #10

Analyzing Evidence from Multiple Sources to Support Construct Validation and Score Interpretations

Jay Thomas, ACT, Inc.; Thomas Langenfeld, ACT, Inc.

Think-aloud and eye-tracking data were analyzed along with pretest data to provide insights into cognitive processing required for solving tasks involving graphics. The proposed model is that the interaction of the graphical complexity and cognitive skills elicited by Graphic Literacy items. Preliminary findings support the model and resulting score interpretations.

Electronic Board #11

ELA & Math Scores Explained by School and Teacher Characteristics

Kimberly Colvin, University at Albany, SUNY

What percentage of the variability in grade 3 to 8 ELA and math scores can be explained by school and teacher characteristics? This study evaluates this question from both the school- and student-levels. The factors investigated include the make-up of the student body and teacher experience and certification.

Electronic Board #12

Feature Aggregations in Automated Essay Scoring

Jing Chen, Human Resources and Research Organization; Mo Zhang, Educational Testing Service; Isaac Bejar, Educational Testing Service

To establish the validity of scores generated by automated scoring systems, the extraction and combination of essay features need to be done in a substantively and technically defensible way. We compare two statistical models that weight and combine essay features to provide recommendations for feature aggregations in automated essay scoring.

Electronic Board #13

Wrong Answers on Classroom Assessments: Benefits for Learning and Motivation

James McMillan, Virginia Commonwealth Universityy

A model of the process students experience when they are wrong on classroom assessments of learning is presented, supported theoretically by research in several areas and some preliminary research on student perceptions of making mistakes and errors, with implications for assessment that will cultivate the positive effects of being wrong.

Electronic Board #14

Embedding Pretest Items in Shadow Tests for Online Calibration

Jie Li, McGraw-Hill Education

This study addresses the practical issues of embedding pretest items in shadow tests for online calibration. These issues include concurrent selection of operational and pretest items, stimuli with pretest items, and infeasibility handling due to the dynamic nature of pretest item pool. A solution is demonstrated through an empirical example.

Electronic Board #15

Pseudo-Simulation on Multistage Test (MST) Design and Analysis Issues

Meng Wu, Educational Testing Services

Analyses of the NAEP MST field trial indicated that the restriction of proficiency range in combination with the limited per student test information at the subscale level interfered with estimating unbiased parameters. This paper will explore alternative data collection designs to address this issue via a pseudo-simulation study.

Electronic Board #16

Matrix-sampling: An effective and efficient approach to assess non-cognitive outcomes

Jesse Pace, University of Kansas; Jesse Pace, University of Kansas; Richard Carter, University of Kansas; Mary Rice, University of Kansas; John Poggio, University or Kansas

Evaluated was the adequacy of matrix-sampling techniques for approximating full-item and group-level parameters of non-cognitive measures. Results indicated that matrix-sampling can approximate parameters with substantially fewer items, and is superior to simply shortening scales. We discuss future research and the need to increase utilization of matrix-sampling for non-cognitive assessments.

Electronic Board #17

Longitudinal Changes in the Relationship Between Test Scores and High School Graduation

David Braslow, Harvard Graduate School of Education

Improvements in state test scores are used to measure improvements in educational quality, but gains may not generalize to later outcomes. I describe changes in the between- and within-school relationships between middle school test scores and high school graduation. Score gains are shown to be spurious, particularly for low-performers.

Electronic Board #18

Invariance of Weighting Schemes in Forming Composite Scores

Qing Xie, University of Iowa/ ACT, Inc; Yi-Fang Wu, ACT, Inc; Xiaohong Gao, ACT, Inc

This study investigates invariance of weighting schemes in forming composite scores across different populations and parallel test forms. Multivariate generalizability theory approaches are applied to real and simulated data. Factors of interest include weighting schemes, population distributions, and pairwise correlations between individual tests.

Electronic Board #19

Asymptotic Sampling Variability for Dependability Coefficient for Single Facet Design

Rashid Almehrizi, Sultan Qaboos University

Estimates of the dependability coefficient are subject to sampling variability. The paper proposed an asymptotic sampling variance for dependability coefficient with delta method. A simulation is performed for normal and nonormal data with different test conditions. Results showed that sampling variance is approaching their true scores.

Electronic Board #20

Assessing Effects of Time Constraints on Examinee Performance on a Licensing Examination

Polina Harik, National Board of Medical Examiners; Brian Clauser, National Board of Medical Examiners; Irina Grabovsky, National Board of Medical Examiners; S Bucak, National Board of Medical Examiners; Michael Jodoin, National Board of Medical Examines; William Walsh, National Board of Medical Examiners; Steven Haist, National Board of Medical Examiners

This paper describes an experiment designed to measure the effects of time limits on examinee performance in the context of a clinical knowledge assessment administered under high stakes. Computerized administration allowed to examine how the response time and the probability of a correct response vary across experimental conditions.

Electronic Board #21

Lognormal Continuous Item Response Models for Multidimensional Compositional Items

Chia-Wen Chen, The Education University of Hong Kong; Wen-Chung Wang, The Education University of Hong Kong

Compositional items yield continuous and ipsative responses. No IRT models were available for this item format. We developed a lognormal compositional Rasch model for compositional items measuring multiple latent traits, and conducted simulations to evaluate its parameter recovery. The results demonstrate a good recovery using joint maximum likelihood estimation methods.

Electronic Board #22

Using Multidimensional Models to Assess the Dimensionality of Technology-Enhanced Items

Ki Matlock, Oklahoma State University; Jonathan Templin, University of Kansas

The purpose of this study is to investigate the dimensional structure of technology-enhanced items in a large-scale testing program using uni- and multidimensional item response theory models and a bifactor diagnostic classification model. Results include a comparison of model-fit, item parameters, and trait score estimates.

Electronic Board #23

A comparison of the performance of computerized adaptive testing and multistage testing

Keyin Wang, Michigan State University

Various CAT and MST designs have been compared under the same item pool (designed for either CAT or MST). This comparison might be unfair to the other testing mode. This study compared the performances of MST and CAT, matched on properties under item pools designed for MST and CAT, respectively.

Sunday, April 30, 2017
2:15 PM–3:45 PM, Salon A, Meeting Room Level, Coordinated Session, L1

Opportunity to Learn: Impact on Large-Scale and Classroom Assessment Design and Interpretation

Session Chair: Kristen Huff, Curriculum Associates
Session Discussant: Steve Ferrara, Consultant

A primary assumption of standards-based assessment is that students have had adequate opportunity to learn the standards on which they are being assessed. When trends are observed that indicate low achievement by one or more subgroups on one or more learning standards, this is considered evidence that educators need additional support in providing effective instruction on those standards. Similarly, in regard to individual student performance, low achievement on particular standards is an indicator of where the student needs additional support. However, an alternative explanation for low performance is that the tasks do not target the learning standard in a way that is sensitive to instruction. When new learning standards are implemented, how to interpret results from assessments of the new standards can be challenging if not confusing. This is especially true as educators, test designers, and policy-makers have various notions of how to interpret learning standards in different contexts (e.g., instruction versus assessment). As such, there is a complex relationship among learning standards, task design, instructional sensitivity, opportunity to learn, and interpretation of results. The nature of this complex relationship and how it plays out in design and score interpretation from both large-scale summative and classroom-based assessments will be addressed.

Item & Test Design Considering Opportunity to Learn and Instructional Sensitivity
Michael Rodriguez, University of Minnesota

Consequences of Ignoring Opportunity to Learn for the Psychometric Quality of Assessments
Paul Nichols, ACT; Pamela Paek, ACT; Britte Cheng, SRI

Principled Assessment Design to Support Formative Assessment and Students' Opportunities to Learn
Christina Schneider, NWEA

The Appropriateness of Off-grade Items in Adaptive, Diagnostic Testing
Shuqin Tao, Curriculum Associates

Sunday, April 30, 2017
2:15 PM–3:45 PM, Conference Room 1&2, Meeting Room Level, Coordinated Session, L2

Understanding, Validating, and Evaluating Assessment: Promoting Sound Decisions through Assessment Literacy

Session Chair: Charles DePascale, Center for Assessment (NCIEA)
Session Discussant: Susan Brookhart, Duquesne University

Data-driven decision-making has become "a mantra of educators from the central office to the school, to the classroom" and assessments are a major source of that data and a focus of state and federal accountability policies (Marsh, Pane, and Hamilton, 2006). Using data from assessments effectively to make sound decisions requires knowledge and skills that are not well-defined and are not sufficiently addressed in the preparation and development of teachers, administrators, and policy makers. This session begins by establishing the importance of assessment literacy in promoting sound decision-making and the need for considering context as an essential element in defining, promoting, and sustaining assessment literacy. The second and third presentations provide two examples of the way in which context impacts the assessment literacy needed to make specific decisions regarding interim assessment programs. The second presentation focuses on establishing the validity of interim assessments, which are often administered under less standardized conditions than traditional summative, large-scale assessments. The third presentation focuses on the specific use of interim assessment results to improve teaching and learning. The fourth presentation presents a toolkit designed to promote assessment literacy in educators while supporting them in evaluating the technical quality of locally-developed, common assessments.

Context in Assessment Literacy and Assessment Literacy in Context
Amy Sharp, University of Washington; Kelli Ryan, Kent State University; Charles DePascale, Center for Assessment; Damian Betebenner, Center for Assessment

Evaluating Validity of An Interim Assessment
Yi-Chen Chiang, Indiana University - Bloomington; Brian Gong, Center for Assessment

Can Interim Assessments Support Improved Teaching and Learning?
Rebecca Ellis, Michigan State University

Evaluating Assessment Quality: A Toolkit to Support Educators' Assessment Literacy
Caitlin Byrne, The University of Alabama

Sunday, April 30, 2017
2:15 PM–3:45 PM, Conference Room 3&4, Meeting Room Level, Coordinated Session, L3

ELPA21: Standard Setting and Score Interpretation

Session Organizer: Daniel Lewis, Pacific Metrics Corporation
Session Chair: Mary Seburn, Quantiful, LLC
Session Discussant: Gregory Cizek, University of North Carolina

English Language Proficiency Assessment for the 21st Century (ELPA21) is a consortium of states that developed and administered for the first time in 2016, an innovative English language proficiency (ELP) assessment. This assessment, ELPA21, measures the English language skills necessary for students to engage and succeed in grade-level academic language demands embedded in the Common Core State Standards (CCSS) and the Next Generation Science Standards (NGSS) at a level comparable to their non-EL peers. Like other ELP assessments, ELPA21 reports skills on four domains—Reading, Writing, Speaking and Listening, but uses profiles, or patterns of performance across the four domains, to determine proficiency rather than rely upon a single composite score. Doing so required innovation in standard setting and psychometric methods, and required multiple sources of data to identify and validate the profiles associated with different levels of proficiency. This coordinated session describes some of those innovations, including new technology implemented for standard setting, alternate ways of writing Achievement Level Descriptors (ALDs), and new measurement models and analytic approaches. The policy and instructional implications of these innovations are discussed by ELPA21's Lead State.

A Profile-Based Approach to English Language Proficiency Assessment
Mary Seburn, Quantiful, LLC.

Design and Technology Innovations for the ELPA21 Bookmark Standard Setting
Daniel Lewis, Pacific Metrics Corporation

Developing Grade-Level ALDs for Grade-Band ELPA21 Assessments
Karla Egan, EdMetric LLC

Incorporating Multiple Standard Setting Methods to Establish Cut Scores for ELPA21
Mark Hansen, University of California, Los Angeles & CRESST.

Proficiency is not a number: One state's transition to profile-based proficiency determinations
Steve Slater, Oregon Department of Education

Sunday, April 30, 2017
2:15 PM–3:45 PM, Salon B, Meeting Room Level, Coordinated Session, L4

Not-your-average mode study: NAEP goes tablet

Session Organizer: Andreas Oranje, ETS
Session Discussant: Matthew Johnson, Teacher's College, Columbia University

The transition of the National Assessment of Educational Progress from paper to (multi-stage) tablet based administration has been widely discussed. What has not been shared yet is the underlying design and statistical machinery that was and is involved with this complicated transition. We have learned a lot of important lessons that, if heeded, will give other assessments a significant head start as they eventually will have to make that change and maintain relevance in a digitized society. For example, we have compared two approaches for calibrating tablet assessment data onto existing (paper originated) scales, each of which rely in different ways on equivalent samples and not-always-common items. We have also learned how multi-stage testing can fail unexpectedly and when it does work well. In addition, we want to share some critical findings about population definition and identification when dealing with non-random calibration designs. All of these studies and findings form the backdrop to a session that will present best-in-class mode transition designs for group score assessments while giving an unprecedented look inside the NAEP psychometric kitchen. Empirical mode study results will be presented based on 2015 national samples in Reading and Mathematics, grades 4 and 8.

Design Principles and Analysis Methodology for Mode Transitions of Group Score Assessments
Andreas Oranje, ETS; Rebecca Moran, ETS

Comparison of Trend Calibration Approaches Across Modes with (Un)Common Items and Random
Xueli Xu, ETS; Paul Jewsbury, ETS

Good and Not-so-good Data Collection Designs for MST in Group-Score Assessments
Meng Wu, ETS; Nuo Xi, ETS; Adrienne Sgammato, ETS

Population Definition and Identification, Priors, and Non-Random Samples
Yue (Helena) Jia, ETS; Meng Wu, ETS; Ru Lu, ETS

Sunday, April 30, 2017
2:15 PM–3:45 PM, Salon C, Meeting Room Level, Coordinated Session, L5

The alignment onion: Multiple layers to validity considerations

Session Discussant: Yvette Nemeth, HumRRO

The reauthorization of ESEA in the form of ESSA has raised many issues related to alignment. In ESSA, alignment is not solely between assessment items and standards but also includes achievement standards to content standards, alternate achievement standards to achievement standards, content standards to college and career readiness standards, and English language proficiency (ELP) standards to content standards. With multiple components needing to show alignment for the validity of a state's assessment system, this session will focus on three areas. First, the alignment of items used in a computer adaptive testing (CAT) environment requires new evaluation criteria. Presenters will discuss alignment results from a state CAT system and Smarter Balanced summative assessments. Second, alignment of achievement standards as an outcome of standard setting is crucial to inferences being made from a student's score indicating what a student knows and can do. Research focused on the connection between items and achievement level descriptors via standard setting will be presented. Third, alignment needs to exist between content standards and ELP standards to ensure that all students are being assessed equally and score inferences are comparable. Methods that may be applied to support the validity of ELP standards and assessments will be presented.

An Innovative Alignment Approach to Computer Adaptive Tests
Hillary Michaels, HumRRO; Sheila Schultz, HumRRO

Validity Evidence in Alignment for Computer Adaptive Tests
Liru Zhang, Delaware Department of Education

Item Mapping and Achievement Level Descriptor Validation
Marty McCall, Smarter Balanced

Alignment is in the Eye of the Beholder: Validity Considerations
Gary Cook, Wisconsin Center for Education Research

Sunday, April 30, 2017
2:15 PM–3:45 PM, Salon D, Meeting Room Level, Paper Session, L6

Structural Equating Modeling

A Latent State-Trait Theory Approach to Deriving Reliability Coefficients for Congeneric Measures
Carrie Morris, University of Iowa; Walter Vispoel, University of Iowa; Murat Kilinc, University of Iowa

A new approach to estimating reliability for congeneric measures based on latent state-trait theory is illustrated that takes multiple sources of measurement error into account. Results reveal that this approach provides higher and more appropriate estimates for reliability than does Generalizability Theory when assumptions of essential tau equivalence are violated.

Constructing Math Motivation Factors and Testing Measurement Invariance
Sakiko Ikoma, American Institutes for Research; George Bohrnstedt, American Institutes for Research; Jizhi Zhang, American Institutes for Research; Bitnara Park, American Institutes for Research; Burhan Ogut, American Institutes for Research; Markus Broer, American Institutes for Research

We explored the dimensionality and stability of the math motivation items and sub-constructs used in the HSLS study and assessed their validity as a part of a larger study on the role of motivation for mathematics performance using the NAEP-HSLS overlap same. The analyses include EFA, CFA and SEM modeling.

Examining the Growth and Achievement of STEM Majors Using Latent Growth Models
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 examined the use of latent growth models (LGM) when investigating the growth and college readiness of STEM majors versus non-STEM majors. Specifically, LGMs were used to compare growth on a state achievement test from Grades 6-11 of STEM majors and non-STEM majors at a public university.

Examining the Relation Between Kindergarten Entry and Emerging Literacy and Math Achievement
Phillip Irvin, University of Oregon; Gerald Tindal, University of Oregon; Steve Slater, Oregon Department of Education

This study examines the predictive-concordant relation between academic and behavioral kindergarten entry skills and spring literacy and math achievement, while accounting for student-level demographics. After confirming the factor structure of a statewide kindergarten entry assessment, we use structural equation modeling to examine the effects of entering skills on later achievement.

Monte Carlo Estimation of the Polychoric Correlation Asymptotic Covariance Matrix
Scott Monroe, University of Massachusetts Amherst

In SEM, ordinal variable models are frequently fit using multistage estimation, which relies on polychoric correlations. Various quantities (e.g., overall test statistics) depend on the asymptotic covariance matrix (ACM) of the polychoric correlations. This research proposes a Monte Carlo approach for computing the ACM, which leads to more accurate estimation.

Sunday, April 30, 2017
2:15 PM–3:45 PM, Salon K, Meeting Room Level, Paper Session, L7

Computerized Adaptive Testing

Session Chair: Xiao Luo, National Council of State Boards of Nursing

An evaluation of content coverage of a computerized-adaptive language test

Kyoko Ito, Defense Manpower Data Center; Tia Sukin, Pacific Metrics; Lihua Yao, Defense Manpower Data Center

Through dimensionality analysis, examinations of computer-adaptive testing (CAT) pools, and evaluations of CAT tests and scores from simulation, the study assessed the impact of not including content constraints in the CAT algorithm for a large-scale language assessment.

Constraint Management in Adaptive Testing with Mixed Format Using R Package CCAT

Qi Diao, Pacific Metrics Corporation; Hao Ren, Pacific Metrics Corporation

Constraint management methods for computerized adaptive testing (CAT) are discussed with mixed format requirements. The methods evaluated are weighted deviation method, maximum priority index, and shadow test approach. R package CCAT was developed to implement these methods in CAT. Empirical examples will be given.

Impact of Item Preknowledge on Measurement Outcomes in Computerized Adaptive Testing

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

The impact of item preknowledge (IPK) on ability estimation was examined analytically through step size and empirically through simulation. IPK was simulated varying selection algorithms, pool characteristics, and location within a fixed-length computerized adaptive test (CAT). Bias, test overlap, and item pool usage were compared between the utilized selection algorithms.

Item Selection Methods in Multidimensional Computerized Adaptive Testing Adopting Polytomously-scored Items

YUTING HAN, School of Psychology, Jiangxi Normal University; DONGBO TU, School of Psychology, Jiangxi Normal University; YAN CAI, School of Psychology, Jiangxi Normal University

This paper introduced the item selection algorithms in Multidimensional Computerized Adaptive Testing (MCAT) with polytomously-scored items. Simulation study was conducted to demonstrate the feasibility of the polytomous MCAT (PMCAT) and to compare the performance among the proposed methods.

New Stopping Rule of Computerized Adaptive Test

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

New CAT stopping rule that incorporates projected future events are proposed and compared with the 95% confidence interval rule in a simulation study and an empirical analysis. Results imply that new stopping rules can further improve in measurement efficiency as they effectively confine the possibilities of future items and responses.

Sunday, April 30, 2017
2:15 PM–3:45 PM, Salon L, Meeting Room Level, Paper Session, L8

Evaluating Model-data Fit in IRT

Session Chair: John Donoghue, Educational Testing Service

A Rotatable Asymmetric Variable Compensation MIRT Model and Its Estimation

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

The purpose of this study is to develop and estimate a variable compensation multidimensional item response theory (MIRT) model that allows for transformation between different correlation structures. The great flexibility and estimability showed by the model allows for it to be used in practice.

Assessing Fit of Unidimensional IRT Models Using KS Statistics

Ying Lu, Educational Testing Service

This study examines the usefulness of Kolmogorove Smirnov (KS) statistics for assessing fit of unidimensional IRT models. The statistics look at the agreement between the expected and the observed examinee ability frequency distribution conditional on each possible item response.

Can Subscores be Detected when the Data Fit Unidimensional Models?

Yi-Ling Cheng, Michigan State University; Mark Reckase, Michigan State University

The current study tested whether it is possible to detect atypical sub-scores from subgroups when the overall dataset fits a unidimensional model well. To explore this problem, we created simulated datasets and then used person fit indices to test whether the samples that had atypical sub-scores could be identified.

Comparison of the Original and Modified S-X2/S-G2 Measures of Item Fit

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

Orlando and Thissen's (2000) S-X2 and S-G2 are among the better functioning measures of IRT item fit. Simulation compared the original (including the studied item in sum score) and modified (removing the studied item) versions of the statistics. Differing degrees of pooling cells with small expected values was also examined.

Evaluating and Diagnosing Item Fit with Tukey-Hann Functions and RISE Statistics

Jeremy Jennings, University of Georgia; George Engelhard, University of Georgia

A framework for exploring item fit using Tukey-Hann Functions (THFs) and RISE statistics is described. The RISE statistics are used to examine the differences between the THFs and model-based IRFs. Simulated data are used to explore Type I error rates and the power of this framework to identify item misfit.

Sunday, April 30, 2017
2:15 PM–3:45 PM, Salon M, Meeting Room Level, Paper Session, L9

Automated Scoring

Sessions Chairs: Andre Rupp, Educational Testing Service; Syed Latifi, University of Alberta

Extreme Scoring Machine: Integrating Deep Language Features for Developing Essay Scoring Framework
Syed Latifi, University of Alberta; Mark Gierl, University of Alberta

We proposed new discourse processing methods for modeling the deep language features for developing three-staged automated essay scoring (AES) framework. Essays from large-scale tests were used. The results outperformed the studied state-of-the-art commercial AES systems and suggested the value and effectiveness of deep features for the development of scoring framework.

Implementing a Contributory Human-machine Scoring Approach for a Large-scale Writing Assessment
Andre Rupp, Educational Testing Service; Jay Breyer, Educational Testing Service; Brent Bridgeman, Educational Testing Service

We present comprehensive empirical evidence for critical methodological design decisions to support the use of a contributory scoring approach for a large-scale writing assessment. We use data from six samples collected during operational administrations and special validity studies and discuss how the methodology is generalizable to other use contexts.

Improving Selection of Automatic Speech Scoring Models
Han Hui Por, Educational Testing Service; Anastassia Loukina, Educational Testing Service

We apply the criticality analysis approach to select scoring models in the automated scoring of spoken responses in a language assessment. We show that this approach addresses issues of sample dependence and bias, and identifies salient features that are critical in improving model validity.

LSTM Cluster: A Novel Way to Cluster Students' Problem Solving Sequences
Qi Guo, University of Alberta; Ying Cui, University of Alberta

This study proposes a novel approach to analyzing students' log files in technology based assessments. The analyses can help teachers better understand how students arrive at a conclusion and design more effective instructional interventions. A simulation study will be conducted to evaluate the proposed method. Initial results seem promising.

Predicting Writing True Scores in Automated Scoring of Essays
Lili Yao, Educational Testing Service; Shelby Haberman, Educational Testing Service; Mo Zhang, Educational Testing Service

In automated scoring, a basic task is prediction of composite true scores of writing. This study examines a widely applicable prediction approach that combines best linear prediction and penalty function to treat lack of population invariance. Predictors include human scores and computer-generated essay features. Key results are provided for illustration.

Sunday, April 30, 2017
2:15 PM–3:45 PM, Salon J, Meeting Room Level, Electronic Board Session : GSIC Graduate Student Poster Session, L10

Graduate Student Issues Committee

Brian Leventhal, Chair

Evelyn Johnson; Dubravka Svetina; Abeer Alamri; Maria Bertling; Brittany Flanery Crawford; David Martinez Alpizar; Rich Nieto

Electronic Board #1

Response Styles as a Source of Person Misfit

Daniel Adams, University of Wisconsin-Madison; Daniel Bolt, University of Wisconsin-Madison

We investigate response style as sources of person fit by applying posterior predictive checks with two discrepancy statistics and applying a multidimensional nominal response model. The methods are studied by simulation and in application to PISA 2006 data.

Electronic Board #2

The Impact of Missingness to the G-DINA Model

Yan Sun, Rutgers, The State University of New Jersey; Yu Bai, Teachers College, Columbia University; Jimmy de la Torre, The University of Hong Kong

Missing responses happen when an examinee misses some items which results in an incomplete data. In this study, the impact of non-ignorable missingness to a general cognitive diagnostic model, the G-DINA model, was detected in terms of the effects on both classification rate and parameter estimation.

Electronic Board #3

A General Procedure for Constructing Intervals for Various Psychometric Indices Using MCMC

Kyung Yong Kim, University of Iowa; Won-Chan Lee, University of Iowa

This study introduces a ``general" interval estimation procedure using MCMC, which can be used to construct intervals for various psychometric indices based on IRT. The interval estimation procedure will be illustrated for constructing intervals for conditional standard error of measurements and classification consistency using real data sets.

Electronic Board #4

Bootstrap Likelihood Ratio Test for Person-Fit under DINA Framework

Yu Bai, Teachers College, Columbia University; Kevin Santos, University of the Philippines; Jimmy de la Torre, The University of Hong Kong

Likelihood Ratio Test (LRT) is a popular person-fit test for identifying examinee's aberrant behavior. LRT is based on either marginalized or joint likelihood. This study proposed to bootstrap LRT statistics for both LRT based on marginalized and joint likelihood to improve the accuracy of a person-fit test.

Electronic Board #5

The impact of DIF methodology on metric identification in Mixture IRT Models

Juyeon Lee, Unversity of georgia; Allan Cohen, University of Georgia

Multiple methods exist for detection of DIF. In this study, we examine the impact of two DIF detection methods on constructing a common metric among latent classes in mixture IRT models. Sample size ratios between latent classes will be examined for their effect on DIF detection and linking accuracy.

Electronic Board #6

Assessing Classification Accuracy and Consistency in a Testlet-Based Test Using IRT

Seohee Park, University of Iowa; Timothy Ansley, University of Iowa

This study uses IRT to investigate Classification Accuracy and Consistency (CA/CC) in a testlet-based test with the Rudner (2001) and Lee (2010) approaches. Various IRT models related with a testlet-based test are considered, and the effect of location of cut scores and test length on CA/CC are also examined.

Electronic Board #7

Subscoring in variable-length computerized adaptive test battery

Jiahui Zhang, Michigan State University; Wei He, NWEA

This study explored the use of collateral information in a variable-length computerized adaptive test battery with the aim to shorten the test length while maintaining the subscore precision. Additionally, a two-stage design was proposed and demonstrated to successfully further shorten test length when different dimensions are highly correlated.

Electronic Board #8

Impact of Mixed Types of Missing data on uniform DIF Detection

Qianqian Pan, University of Kansas; Wei Wu, University of Kansas

A simulation study was conducted to investigate the impact of mixed types of missing data on uniform differential item function (DIF) detection. This study extended earlier work by including missing responses on both DIF and DIF-free items. The result suggests that missingness on DIF-free items can impact DIF detection.

Electronic Board #9

A Comparison of UIRT and MIRT equating methods

Youkyoung OH, Yonsei University

The purpose of this study is comparing the adequacy of multidimensional and unidimensional IRT equating methods under the random groups design in particular with mixed-format tests by extending the work of Lee and Brossman (2012). It is expected we can provide specific guidelines for equating studies with the mixed-format test.

Electronic Board #10

Application of Hybrid Computerized Adaptive Testing

Yutong Wang, Beijing Normal University

Computerized adaptive testing (CAT) and Multistage Testing (MST) have become two of the most popular trends in large scale assessment. The hybrid computerized adaptive test combined MST and CAT, setting the test more easer. The application of the hybrid test is designed, and getting a good result.

Electronic Board #11

IRT-based models as a proxy to the t-test to detect group difference

Saemi Park, The Ohio State University; Paul de Boeck, The Ohio State University

We compare (1) t-tests of group effects on sumscores of binary data and (2) IRT-based tests for the same group effects. The latter have more power and a more inflated Type I error, and IRT-based correlations of group with the dependent variable are larger than correlations with the sum score.

Electronic Board #12

Measurement Invariance of Transadapted Assessments Using Three Psychometric Frameworks.

Duy Pham, University of Massachussets Amherst

This study adopts classical test theory, item response theory and structural equating modeling frameworks to investigate measurement invariance of a translated assessment. Preliminary findings support partial construct equivalence across lingual forms. At item level, differential item functioning results are quite consistent across three detection methods.

Electronic Board #13

A Review and Comparison of Polyserial Correlation Indices

Nathan Minchen, Rutgers, The State University of New Jersey; Ying Lu, Educational Testing Service

Item discrimination is an important index to consider in test construction. Many tests continue to use traditional methods of calculating item discrimination, which are based on item-total correlations. Our study compares several commonly used coefficients, one of which is a lesser-known but highly accurate coefficient that has computational advantages.

Electronic Board #14

Bayesian Framework for Estimation of a General Class of Cognitive Diagnostic Models

Susu Zhang, University of Illinois at Urbana-Champaign; Steven Culpepper, University of Illinois at Urbana-Champaign

Based upon a general class of cognitive diagnostic models developed by Xu with minimal model restrictions, we propose a Bayesian parameter estimation algorithm for these models using Gibbs sampling. Simulation studies suggested that the proposed method accurately recovered true item and person parameters and demonstrated high computational efficiency.

Electronic Board #15

Fitting Nominal Response Model with Testlet Items

Anqi Li, University of Illinois, Urbana-Champaign; Hua-Hua Chang, University of Illinois, Urbana-Champaign

This study investigates the performance of nominal response model on both dichotomous and polytomous testlet items. First, its performance is compared with both the generalized partial credit model and graded response model on dichotomous testlet items; and second, an extension is proposed to better fit with polytomous testlet items.

Electronic Board #16

Propensity Score Estimation in Multilevel Setting: Data Mining Methods as Alternatives

HyunSuk Han, University of Florida; Minho Kwak, University of Georgia

The purpose of this study is to compare several propensity score methods including data mining methods with traditional propensity score methods such as logistic and multilevel logistic regression models. Data mining methods such as random forest (RF) and generalized boosted model (GBM) are considered in this study.

Electronic Board #17

Weighted Likelihood Procedures for Ability Estimation in Mixed-Type Tests

Alvaro Cruz, University of Illinois at Urbana-Champaign

This simulation study will look into item-weighted likelihood procedures for ability estimation in tests containing polytomous and dichotomous items. For the former NRM is implemented and for latter the 3PL model. Using Tao & Chang (2012) as guide, estimation ability will be assessed for the weighted procedures against the MLE.

Electronic Board #19

Robust Bayesian Estimation of Item Response Model Parameters Accounting for Aberrance

Kaiwen Man, The university of Maryland at College Park; Hong Jiao, University of Maryland at College Park; Yong Luo, National Center for Assessment in Higher Education, Riyadh, Saudi Arabia; Yunbo Ouyang, University of Illinois at Urbana and Champaign

Aberrant responding behaviors introduce more measurement errors in the estimation of latent parameters in Item Response Theory (IRT) models with maximum-likelihood estimation method.This study explores the robust Bayesian method to recover the "true" estimates and the effectiveness of the different robust priors is evaluated under different simulated study conditions.

Electronic Board #20

Investigation of cheating during breaks on a large-scale regulatory exam

Yating Zheng, University of Maryland at College Park; Elena Boiarskaia, FINRA

This study explores three methods to detect potential cheating during breaks on a large-scale LOFT exam. These methods focus on answer changes before and after breaks using response time and Rasch ability estimates for comparison. This is a novel approach to identify potential use of test preparation materials during breaks.

Electronic Board #21

Person Fit across Subgroups: PISA 2012 Mathematics Assessment

Mingying Zheng, University of Nebraska-Lincoln

Person fit is often used to evaluate model-data fit. Investigating misfitting item score patterns across different groups is strongly related to differential item functioning (DIF). In this study, the person fit and item score patterns for different groups were investigated using PISA 2012 mathematics assessment (USA).

Electronic Board #22

Mixture Item Response Model with Covariates in Detection of Latent DIF

Ya Zhang, University of Pittsburgh

A simulation study is performed to investigate the effectiveness of mixture IRT models with covariates (MixIRT-C) in DIF analysis. The effect size of DIF, association between covariates and latent classes, and distribution of latent traits are manipulated. The study provides comprehensive evaluation of the performance of MixIRT-C models.

Electronic Board #23

Different Inferences Based on the Type of Propensity Score Matching Method Used

Alejandra Garcia, University of Massachusetts - Amherst

Propensity scores are used to match groups in studies where participant randomization is not feasible. Many propensity score matching (PSM) methods are currently used. This study applies multiple PSM methods to an educational data set to determine if, when covariates are matched satisfactorily, the PSM method affects the data interpretation.

Sunday, April 30, 2017
4:00 PM–7:00 PM, Conference Room 10, Meeting Room Level

NCME Board of Directors Meeting

Sunday, April 30, 2017
4:05 PM–6:05 PM, Salon A, Meeting Room Level, Invited Session, M1

Diversity Issues in Testing Committee Sponsored Symposium

Session Chair: Lietta Scott, Arizona Department of Education
Session Panelists: Martha Thurlow, National Center on Educational Outcomes
Rachel Kachchaf, Smarter Balanced Assessment Consortium
Nathan Wall, eMetric, San Antonio
Stephanie Cawthon, University of Texas at Austin
Laurene Christensen, WIDA
Cara Laitusis, Educational Testing Service

With the new advancements in innovative technology and computerized testing, both in large-scale and classroom assessments, it is crucial to not only foster connections between standardized large-scale testing and classroom use, but also ensure that the summative and formative assessments are accessible and designed with diverse test-takers in mind. In this session, invited accessibility experts will discuss the importance of accommodation and accessibility features in large-scale and classroom assessments, and their impact on the validity of the score interpretations made from these tests. This session will be structured as a moderated panel-discussion where experts will respond to directed questions from various perspectives (including specific accommodations provided for English Language Learners and Students with Disabilities). Topics focus both on the importance of accessibility features in designing parallel forms and on the specific validity concerns that impact score interpretations for diverse test-takers when large-scale and classroom assessments are designed with and without accommodations. The discussions in this session will be particularly useful to practitioners and might help guide end-users and stakeholders with specific recommendations in considering accessibility features at every stage of test design and development as well as providing better guidance on the appropriate interpretation and usage of test scores.

Sunday, April 30, 2017
4:05 PM–6:05 PM, Conference Room 1&2, Meeting Room Level, Coordinated Session, M2

Assessment in MOOCs: Current and Next Generation Research and Development

Session Organizer: Yigal Rosen, Harvard University
Session Chair: Daniel Seaton, Harvard University
Session Discussants: Steve Ferrara, Measured Progress; Yoav Bergner, New York University

Assessment in Massive Open Online Courses (MOOCs) encompasses broad definitions and endless possibilities. In this symposium, we first summarize the fundamental types of assessment used broadly in the edX platform through programs at Harvard University and Davidson College. Harvard's MOOC organization offers a glimpse at a large scale course provider with tremendous topical diversity, while Davidson College represents a small, targeted program aimed at Advanced Placement courses. We are loosely framing this proposal as a past, present, future look at assessment in MOOCs. The submission will start with an overview of MOOC assessment from the perspective of a couple institutions (Harvard, MIT, Davidson, and ETS). Each session should attempt to address what has been done at their institutions and where they want to take assessment in the future.

Assessment Across Multiple Models of AP High School Instruction with MOOCs
Daniel Seaton, Harvard Universtiy

Measuring Assessment Authenticity in Open Online Learning
Curtis Northcutt, MIT; Tailin Wu, MIT; Martin Segado, MIT; Isaac Chuang, MIT

Toward Future HarvardX Assessment
Yigal Rosen, Harvard University

Exploring the scalability of conversation-based assessments in multiple science domains
Lei Liu, Education Testing Service; Jonathan Steinberg, ETS; Farah Qureshi, ETS; Isaac Bejar, ETS; Fred Yan, ETS

Assessment Across Multiple Models of AP High School Instruction with MOOCs
John Hansen, Harvard Universtiy; Goff Julie, Davidson College; Patrick Sellers, Davidson College; Aaron Houck, Queens University of Charlotte

Sunday, April 30, 2017
4:05 PM–6:05 PM, Conference Room 3&4, Meeting Room Level, Coordinated Session, M3

Designing and Evaluating Score Reports for Specific Audiences

Session Chairs: Priya Kannan, Educational Testing Service; Sharon Slater, Educational Testing Service
Session Discussant: April Zenisky, University of Massachusetts, Amherst

Score reports are often the only point of interaction between some score users (e.g., parents) and the assessment. Moreover, different score users (e.g., test-takers, parents, and teachers) often have different levels of familiarity with not only the assessment but also the psychometrics behind the scores reported. Therefore, they often have trouble understanding the information presented in reports. Several strategies have been recommended to help users understand and draw reasonable conclusions from the information presented in score reports. These include taking into account the specific characteristics of the target audience (Zapata-Rivera & Katz, 2014) and best practices (Goodman & Hambleton, 2004) in iteratively designing score reports (Hambleton & Zenisky, 2013) which are vetted with experts and evaluated with the target audience at every step of the process. The papers in this session focus on the various stages of the iterative score report design and development process (i.e., audience analysis – prototype development – evaluation of usability and comprehension – on-going maintenance) for a variety of audiences (i.e., students, parents, and educators). The collection of studies in this session highlight the importance of the iterative method to the validity of interpretations made from assessment results.

Parent Perspectives on Summative Score Reports
Jacqueline Dickey, University of Massachusetts, Amherst; Priya Kannan, Educational Testing Service; Francis Rick, University of Massachusetts, Amherst; Stephen Sireci, University of Massachusetts, Amherst

Evaluating Parent Comprehension of Measurement Error Presented in Score Reports
Priya Kannan, Educational Testing Service; Andrew Bryant, Educational Testing Service; Diego Zapata-Rivera, Educational Testing Service; Stephanie Peters, Educational Testing Service

Designing Effective Score Reports for English Language Learners
Alexis Lopez, Educational Testing Service; Jonathan Schmidgall, Educational Testing Service; Ian Blood, Educational Testing Service; Jennifer Wain, Educational Testing Service

Before and After: Changes to Supplementary Feedback Reports in Medical Licensure
Liane Patsula, Medical Council of Canada; Andre De Champlain, Medical Council of Canada; Andrea Gotzmann, Medical Council of Canada

Educator Interpretation and Application of Assessment Results in a Dynamic Reporting System
Linette McJunkin, Computerized Assessments & Learning; Sharon Slater, Educational Testing Service

Sunday, April 30, 2017
4:05 PM–6:05 PM, Salon B, Meeting Room Level, Coordinated Session, M4

Device effects research in NAPLAN transition to online testing

Session Organizer: Laurie Davis, Pearson
Session Chair: Stanley Rabinowitz, Australian Curriculum, Assessment and Reporting Authority (ACARA)
Session Discussant: Walter Way, Pearson

In preparation for the 2017 implementation of online adaptive testing for the National Assessment Program program—Literacy and Numeracy (NAPLAN), the Australian Curriculum, Assessment and Reporting Authority (ACARA) initiated a comprehensive research and development program which included research to evaluate the effects of device on student testing experience and test score. This session discusses results from two phases of device effects research—the first conducted within a linear test administration format and the second conducted within an adaptive test administration format. In the first phase of research, more than 3500 students at four grade levels from across Australia participated in two one-hour study sessions to evaluate the impact of different devices (PC, tablet, and tablet with external keyboard) on student testing experience and test performance. Qualitative and quantitative results are reported. In the second phase of research, a nationally representative sample of students (2400 students at each of four grade levels) responded to NAPLAN multi-stage adaptive tests, administered on a range of devices. A separate study investigated the impact of testing device on the comparability of results between adaptive online and paper linear NAPLAN tests. The importance of device familiarity in supporting claims of score validity and comparability is discussed.

Qualitative Analysis of Student Device Use in NAPLAN
Laurie Davis, Pearson; Anna Cohen, Australian Curriculum, Assessment and Reporting Authority (ACARA); Kay Treacy, Australian Curriculum, Assessment and Reporting Authority (ACARA); Irene Janiszweska, Pearson; Laura Holland, Pearson

Rasch Analysis of Device Effects in NAPLAN
Robert Schwartz, Pearson; Ben Ben Businovski, Australian Curriculum, Assessment and Reporting Authority (ACARA); Laurie Davis, Pearson

Application of Generalized Linear Mixed Modeling to Evaluate Device Effects in NAPLAN
Goran Lazendic, Australian Curriculum, Assessment and Reporting Authority (ACARA); Robert Schwartz, Pearson; Laurie Davis, Pearson

Analysis of Device Effects in the Context of Multistage Adaptive NAPLAN Tests
Stanley Rabinowitz, Australian Curriculum, Assessment and Reporting Authority (ACARA); Goran Lazendic, Australian Curriculum, Assessment and Reporting Authority (ACARA)

Mode Effects between Paper and Different Digital Devices in NAPLAN
Goran Lazendic, Australian Curriculum, Assessment and Reporting Authority (ACARA)

Sunday, April 30, 2017
4:05 PM–6:05 PM, Salon C, Meeting Room Level, Coordinated Session, M5

Testing, testing: What is the fairest score when applicants retake admissions tests?

Session Chair: Andrew Ho, Harvard Graduate School of Education

Applicants are retaking admissions tests at increasing rates (Harmston & Crouse, 2016), but guidance for appropriate scoring across retakes remains dated (Boldt, Centra, & Courtney, 1986) or limited in scope (Roszkowski & Spreat, 2016). This symposium includes four papers that expand the research base on retaking patterns using recent datasets from the ACT, GRE, and MCAT testing programs. Leveraging future outcomes and changes in test versions, these papers offer evidence-based guidance for scoring while raising questions about fairness when applicants have differential access to retesting opportunities. In Paper 1, Mattern and Radunzel use ACT data to assess differential prediction of first-year grade point averages (FGPA) across retakes. They find that "superscoring" across subject scores maximizes prediction of while minimizing differential prediction. In Paper 2, Bertling and Ho use a sample of ACT data that includes sociodemographic variables. They model college admission and FGPA to show that superscoring is only more equitable if access to retesting is, too. In Paper 3, Ling and colleagues present contrasting findings that show nonsignificant differences in FGPA prediction across GRE scoring methods. Finally, in Paper 4, Jin and Kroopnick use retesting evidence across a change in MCAT version to offer practical guidance to admissions officers.

Differential prediction of college success: The impact of retesting and scoring policies
Krista Mattern, ACT, Inc.; Justine Radunzel, ACT, Inc.

Demographics and differential advantages in college admissions retesting
Maria Bertling, Harvard Graduate School of Education; Andrew Ho, Harvard Graduate School of Education

Investigating validity consequences when using multiple scores on graduate level admission tests
Guangming Ling, Educational Testing Service; David Klieger, Educational Testing Service; Brent Bridgeman, Educational Testing Service; Lydia Liu, Educational Testing Service; Jennifer Bochenek, Educational Testing Service; Chelsea Ezzo, Educational Testing Service

A comprehensive analysis of MCAT retesting gains within and across versions
Ying Jin, Association of American Medical Colleges; Marc Kroopnick, Association of American Medical Colleges

Sunday, April 30, 2017
4:05 PM–6:05 PM, Salon D, Meeting Room Level, Paper Session, M6

DIF Session 3

Session Chair: Xin Luo, Northwest Evaluation Assoication

Detecting Mode Effect Items Using DIF Methods with Different Conditional Variables

Dong-In Kim, Data Recognition Corporation; Christie Plackner, Data Recognition Corporation; Ping Wan, Data Recognition Corporation; Jungnam Kim, National Board of Chiropractic Examiner

To detect mode effect items between online and paper-pencil tests, three popular DIF methods, MH, LR, and IRTAM are compared with different conditional variables from different data sources.Type I error and Power are compared across different degrees and number of mode effect items.

DIF Detection in Polytomous Items: Cross-Classified Modeling of DLID

Wei Xu, University of Florida; David Miller, University of Florida

Relatively few studies addressed DLID when conducting DIF on polytomous items. In this proposed study, we address such research void by proposing a polytomously cross-classified model and investigate the performance of the proposed model using simulated data. The proposed study has great implications for test revision and construction

Measurement Invariance of a Test Battery Taken in Different Orders

Benjamin Andrews, ACT

The measurement invariance of different test orders within a battery is assessed using IRT methods. Item- and test-level information were evaluated for both paper and online administrations. Results showed that some tests functioned differently depending on the position within the battery, particularly for low-performing students.

Methods for Assessing Measurement Invariance for Disparate Sample Sizes and Proficiency Distributions

Craig Wells, University of Massachusetts Amherst; Stephen Sireci, University of Massachusetts Amherst; HyunJoo Jung, University of Massachusetts Amherst

The purpose of this study is to propose and evaluate statistical methods based on two sampling procedures for assessing measurement invariance when the groups being compared differ considerably with respect to sample size and proficiency.

Using a Log-Likelihood-Based Item-Fit Index to Detect DIF Items in CAT

Xin Luo, Northwest Evaluation Association; Mark Reckase, Michigan State University

This study used a log-likelihood-based item-fit index to identify the field-test (FT) items with DIF in a variety of computerized adaptive test (CAT) settings. The results indicated that the lz can achieve high detection power and was promising for future application.

Using Generalized Logistic Regression to Target DIF Item Review across Multiple Groups

Fusun Sahin, University at Albany, State University of New York; Ronli Diakow, New York City Department of Education; Anthony Banners, New York City Department of Education

Generalized logistic regression is a suitable method for evaluating DIF across multiple focal groups with varying sample sizes. Results from a large-scale test offered in five languages showcase the benefits of this approach. Item review is facilitated by controlling over-flagging and targeting reviewers' attention to specific groups in flagged items.

Sunday, April 30, 2017
4:05 PM–6:05 PM, Salon K, Meeting Room Level, Paper Session, M7

Model fit issues with Diagnotic Classification Models

Adjusting Person Fit Index for Skewness in Cognitive Diagnosis Modeling
Kevin Carl Santos, University of the Philippines-Diliman; Jimmy de la Torre, The University of Hong Kong; Matthias von Davier, Educational Testing Service

This study examines different ways of adjusting for skewness of the standardized log-likelihood statistic lZ in the cognitive diagnosis model framework. The performance of the adjusted person fit statistics is investigated using a simulation study. Results show that using the χ2-approximation correction yields better type I error rates and power.

An Empirical Comparison of Two CDMs for Polytomous Attributes
Levent Yakar, Hacettepe University; Jimmy de la Torre, The University of Hong Kong; Wenchao Ma, Rutgers University

This paper proposes the fully additive model (fA-M), and compares it with the polytomous generalized deterministic input, noisy "and" gate (pG-DINA model) using real data. Different statistics indicate that the fA-M provides a better model-data fit than the pG-DINA model. Moreover, the two models produces very disparate examinee classifications.

Applying M2 Statistic to Evaluate the Fit of Hierarchical Diagnostic Classification Models
Fu Chen, Beijing Normal University; Tao Xin, Beijing Normal University; Yanlou Liu, Qufu Normal University

The performance of M2 statistic in evaluating the fit of hierarchical diagnostic classification models was investigated in the presence of 4 types of attribute hierarchies. The findings suggested proper empirical Type I error rates and high statistical powers for M2 under conditions of different sample sizes and attribute correlations.

New Item-level Model Selection Procedures for Diagnostic Classification Models
Tao Xin, Beijing Normal University; Yanlou Liu, Qufu Normal University; Wei Tian, Beijing Normal University; Lingqing Li, Qufu Normal University

In this paper, we proposed a new set of item-level model selection methods based on information estimation procedures for diagnostic classification models. Monte Carlo simulations were conducted to examine the Type I error rate control and power of the new methods. Practical implications are provided.

Testing Model Fit for Non-nested Cognitive Diagnosis Models
Yanhong Bian, Rutgers, the State University of New Jersey; Kevin Carl Santos, University of the Philippines-Dillman; Jimmy de la Torre, The University of Hong Kong

This study extends the Vuong test and the distribution free test to non-nested cognitive diagnosis models (CDMs). The results show that both tests can select the most appropriate non-nested CDM, especially when the item quality is high.

Two-Step Approximations for Model Comparison in Cognitive Diagnosis Modeling
Miguel Sorrel, Universidad Autónoma de Madrid; Jimmy de la Torre, The University of Hong Kong; Francisco Abad, Universidad Autónoma de Madrid; Julio Olea, Universidad Autónoma de Madrid

Selecting one cognitive diagnosis model (CDM) from a set of CDMs at the item level is of critical concern because this decision has implications on classification accuracy. This paper proposes two-step approximations to existing item-level model comparison tests, and investigates the performance of the new procedures using a simulation study.

Sunday, April 30, 2017
4:05 PM–6:05 PM, Salon L, Meeting Room Level, Paper Session, M8

Applications of Validity Methods

Session Chair: Joseph Stevens, University of Oregon

A Reasoned Approach to Considering Testing Stakes

Richard Tannenbaum, Educational Testing Service (ETS); Michael Kane, Educational Testing Service (ETS)

We propose that testing stakes be thought of as a differentiated profile of consequences, rather than as a simply classification of high or low. We provide generalizable criteria for evaluating stakes with applications to licensure, employment, and k-12 accountability testing programs.

Construct Validation of Scenario-Based Language Assessment (SBLA) Tasks for Diagnostic Placement Purposes

Eunice Eunhee Jang, Ontario Institute for Studies in Education, University of Toronto

Based on a multi-round SBLA development and validation project over three year, the present paper discusses the principles of scenario-based language assessment task design, their psychometric rigor necessary for construct validity, and implications for guiding program placement and instruction in a college context.

Integrating validity evidence from response processes analysis, expert interviews and item difficulty

Olga Zlatkin-Troitschanskaia, Johannes Gutenberg-University Mainz; Sebastian Brueckner, Johannes Gutenberg-University Mainz; James Pellegrino, University of Illinois at Chicago

We illustrate how data from an analysis of response processes can be quantitatively integrated with data on item parameters and from expert interviews to validate a test of economic knowledge using multi-level analysis. We found our approach effective for improving the contribution of response process analysis in validation Research.

The Validity of Classroom Observation Systems in Research and Applied Contexts

Shuangshuang Liu, Educational Testing Service; Courtney Bell, Educational Testing Service; Nathan Jones, Boston University

Practitioners often presume that using a previously "validated" instrument correctly will produce "valid" scores. This paper describes the use and preliminary psychometric evidence of a popular teacher observation instrument in two contexts – research and applied. We describe differences between observation systems and discuss how context may shape instrument validation.

Using Effect Size Measures to Estimate and Report Achievement Gaps

Joseph Stevens, University of Oregon; Daniel Anderson, University of Oregon; Joseph Nese, University of Oregon; Gerald Tindal, University of Oregon

We describe common effect size (ES) methods, new or rarely used methods, and ways to interpret and contextualize ES for reporting achievement gaps. We report results in mathematics and reading in three state accountability datasets with emphasis on academic growth and methods to understand group differences across the score distribution.

Validating Quantitative Text Analysis Tools Against Students' Reading Performances: Problems and Solutions

Kathleen Sheehan, Educational Testing Service

New standards increase text complexity expectations for students at all grade-levels. Teachers, publishers and test-developers have been encouraged to consider feedback obtained via quantitative tools when aligning texts selected for use in instruction and assessment. This paper considers whether existing validation approaches are sufficient to support these new, high-stakes applications.

Sunday, April 30, 2017
4:05 PM–6:05 PM, Salon M, Meeting Room Level, Paper Session, M9

Reliability of Ratings

Session Chair: Nicholas Curtis, James Madison University

Agreement between Human and Model-Based Classification Probabilities for a Learning Progression

Edith Aurora Graf, Educational Testing Service; Peter van Rijn, ETS Global; Daniel McCaffrey, Educational Testing Service

Items from a large-scale assessment were scored using learning progression levels. A constrained version of a polytomous IRT model was used to classify students into levels. As a validation exercise, model-based scores were compared to human ratings. We extend this approach by comparing human-produced classification probabilities to model-based classification probabilities.

Detecting Regression Effects in Angoff Ratings Using Standard Deviations

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

One common error panelists often make in Angoff standard setting is that they regress ratings toward the center of the probability scale. This study introduces two indices to detect regression effects in Angoff ratings using ratios of standard deviations and compares these indices to several existing methods.

Exploring Longitudinal Rater Fit in Performance-Based Assessment

Tara McNaughton, Measurement, Inc.

Rating quality as demonstrated by Rasch fit statistics was explored longitudinally within a medical certification exam. A hierarchical linear model (HLM) analysis suggested that experience did not improve fit. The standard deviation of rater fit across years was relatively low for most raters (79%), although some raters demonstrated more variation.

Investigating the Validity of Performance Ratings on a Clinical Skills Licensure Examination

William Roberts, National Board of Osteopathic Medical Examiners; Yuxi Qiu, University of Florida; Larissa Smith, National Board of Osteopathic Medical Examiners

Rater-mediated assessment becomes increasingly part of licensing/credentialing performance examinations. Yet, validity of scores is questioned without objective evidence to support indirect ratings of performance. This study investigates the psychometric quality of experts' ratings of medical students' clinical skill performance using Many-facet Rasch measurement within a complex judging design.

Rubrics: Old News? Using logic-based decision maps to improve performance ratings

Nicholas Curtis, James Madison University; Madison Holzman, James Madison University; Allison Ames, James Madison University

Even with a well-developed rubric, raters of performance assessments must still hold an array of complex information in their mind; raters may still drift; raters may still alter their scores. Thus, we propose a new scoring method grounded in logic-based decision mapping to address these issues.

Participant Index

Participant Index

Participant Index

Participant Index

Participant Index

Participant Index

Participant Index

Participant Index

Participant Index

Participant Index

Participant Index

Participant Index

A

Akande, Christiana
University of Florida
cakande@ufl.edu

Aksu Dunya, Beyza
University of Illinois at Chicago
baksu2@uic.edu

Albano, Anthony
University of Nebraska-Lincoln
albano@unl.edu

Ali, Usama
Educational Testing Service
uali@ets.org

Allen, Jeff
ACT, Inc.
jeff.allen@act.org

Almehrizi, Rashid
Sultan Qaboos University
mehrzi@squ.edu.om

Anderson, Daniel
University of Oregon
daniela@uoregon.edu

Andrade, Heidi
University at Albany-SUNY
handrade@albany.edu

Andrews, Benjamin
ACT
benjamin.andrews@act.org

Ankenmann, Robert
The University of Iowa
robert-ankenmann@uiowa.edu

Arneson, Amy
University of California, Berkeley
amy.arneson@berkeley.edu

Atalmis, Erkan
Kahramanmaras Sutcu Imam University
erkanatalmis@gmail.com

Austin, Bruce
Washington State University
bwaustin@wsu.edu

Ayturk, Ezgi
Fordham University
eayturk@fordham.edu

B

Babcock, Ben
The American Registry of Radiologic Technologists
ben.babcock@arrt.org

Baldwin, Peter
National Board of Medical Examiners
PBaldwin@nbme.org

Bao, Yu
University of Georgia
yubao02@uga.edu

Barrett, Michelle
Pacific Metrics Corporation
mbarrett@pacificmetrics.com

Barry, Carol
College Board
cabarry@collegeboard.org

Barton, Karen
NWEA
karen.barton@nwea.org

Basaraba, Deni
Bethel School District #52
Deni.Basaraba@bethel.k12.or.us

Becker, D. E. (Sunny)
HumRRO
sbecker@humrro.org

Becker, Kirk
Pearson
kirk.becker@pearson.com

Beimers, Jennifer
Pearson
jennifer.beimers@pearson.com

Bejar, Isaac
ETS
ibejar@ets.org

Bell, Courtney
Educational Testing Service (ETS)
cbell@ets.org

B | Contact Information for Individual and Coordinated Sessions First Authors

Belov, Dmitry
Law School Admission Council
dbelov@lsac.org

Benton, Stephen
The IDEA Center
steve@ideaedu.org

Bergner, Yoav
New York University
yoav.bergner@nyu.edu

Bertling, Jonas
Educational Testing Service
jbertling@ets.org

Bertling, Maria
Harvard Graduate School of Education
mbertling@g.harvard.edu

Betts, Joe
Pearson VUE
jbetts5118@aol.com

Beverly, Tanesia
University of Connecticut
tanesia.beverly@uconn.edu

Bian, Yanhong
Rutgers, the State University of New Jersey
hellobyh@gmail.com

Biancarosa, Gina
University of Oregon
ginab@uoregon.edu

Binici, Salih
Florida Department of Education
Salih.Binici@fldoe.org

Bishop, Kyoungwon
University of Wisconsin
gyoungwonlee@hotmail.com

Bishop, Kyoungwon
Wisconsin Center for Education Research – WIDA Consortium
kbishop@wisc.edu

Bolt, Daniel
University of Wisconsin, Madison
dmbolt@wisc.edu

Bonner, Sarah
Hunter College, CUNY
sbonner@hunter.cuny.edu

Braslow, David
Harvard Graduate School of Education
david.braslow@gmail.com

Briggs, Derek
University of Colorado
derek.briggs@colorado.edu

Brocato, Nicole
Wake Forest University
brocatnw@wfu.edu

Brodersen, Alex
University of Notre Dame
abroders@nd.edu

Brookhart, Susan
Brookhart Enterprises LLC
susanbrookhart@bresnan.net

Brown, Gavin
The University of Auckland
gt.brown@auckland.ac.nz

Bruce, Wesley
Independent Educational Consultant
wesbruce3@gmail.com

Buchholz, Janine
German Institute for International Educational Research
Buchholz@dipf.de

Buckendahl, Chad
ACS Ventures, LLC
cbuckendahl@acsventures.com

Buckley, Katie
Transforming Education
buckley.kate@gmail.com

Bulut, Okan
University of Alberta
bulut@ualberta.ca

Bunch, Michael
Measurement Incorporated
mbunch@measinc.com

Burkhardt, Amy
University of Colorado, Boulder
amy.burkhardt@colorado.edu

Burrus, Jeremy
Center for Innovative Assessments, ProExam
jburrus@proexam.org

Byrne, Caitlin
The University of Alabama
cabyrne1@crimson.ua.edu

C

Cai, Li
CRESST/UCLA
lcai@ucla.edu

Cai, Liuhan
University of Nebraska-Lincoln
cliuhan@gmail.com

Canto, Phil
Florida Department of Education
phil.canto@fldoe.org

Cappaert, Kevin
Pearson
cappaer3@uwm.edu

Carlson, Sarah
University of Oregon
carlsons@uoregon.edu

Cartwright, Emmett
Houghton Mifflin Harcourt
emmett.cartwright@hmhco.com

Castellano, Katherine
Educational Testing Service
KEcastellano@ets.org

Chang, Chi
Michigan State University
changc65@msu.edu

Chang, Hua-Hua
University of Illinois at Urbana-Champaign
hhchang@illinois.edu

Chang, Yufeng
Minnesota Department of Education
Yu-Feng.Chang@state.mn.us

Chao, Hsiu-Yi
National Chung Cheng University
hsiuyi1118@gmail.com

Chattergoon, Rajendra
University of Colorado, Boulder
Rajendra.Chattergoon@colorado.edu

Chen, Chia-Wen
The Education University of Hong Kong
s1115795@s.eduhk.hk

Chen, Chi-Chen
National Sun Yat-sen University
runrice@hotmail.com

Chen, Cong
University of Illinois at Urbana–Champaign
cchen105@illinois.edu

Chen, Fu
Beijing Normal University
f.chen@mail.bnu.edu.cn

Chen, Haiqin
American Dental Association
chen.haiqin2010@gmail.com

Chen, Jianshen
Educational Testing Service
jchen006@ets.org

Chen, Jie
University of Kansas
xiaojiewd@hotmail.com

Chen, Jing
Human Resources and Research Organization
jchen@humrro.org

Chen, Jinsong
SYSU
jinsong.chen@live.com

Chen, Juan
National Conference of Bar Examiners
jchen@ncbex.org

Chen, Jyun-Hong
National Sun Yat-sen University
horishana@gmail.com

Chen, Pei-Hua
National Chiao Tung University
peihuamail@gmail.com

Chen, Yi-Hsin
University of South Florida
ychen5@usf.edu

Chen, Yunxiao
Emory University
yx.chen1988@gmail.com

C | Contact Information for Individual and Coordinated Sessions First Authors

Cheng, Yi-Ling
Michigan State University
chengyil@msu.edu

Chiang, Yi-Chen
Indiana University-Bloomington
chiangy@indiana.edu

Chiu, Chia-Yi
Rutgers, The State University of New Jersey
chia-yi.chiu@gse.rutgers.edu

Cho, YoungWoo
ACT Inc.
youngwoo.cho@act.org

Choe, Edison
ACT
e3d2i1@gmail.com

Choi, Hye-Jeong
University of Georgia
hjchoi1@uga.edu

Choi, Jaehwa
The George Washington University
jaechoi@gwu.edu

Choi, Jinah
The University of Iowa
jinah-choi@uiowa.edu

Choi, Jiwon
University of Iowa /ACT, Inc.
jw0326@gmail.com

Choi, Kilchan
CRESST/UCLA
kcchoi@ucla.edu

Choi, Youn-Jeng
University of Georgia
ychoi26@ua.edu

Christensen, Laurene
University of Wisconsin - Madison
laurene.christensen@wisc.edu

Christopherson, Sara
Wisconsin Center for Education Products and Services
sara.christopherson@wceps.org

Chung, Seungwon
University of California, Los Angeles
sxc21@ucla.edu

Cizek, Gregory
University of North Carolina at Chapel Hill
cizek@unc.edu

Clark, Amy
University of Kansas
akclark@ku.edu

Clark, Hope
ACT
Pamela.Paek@act.org

Clarke, Alix
University of Alberta
alix1@ualberta.ca

Clauser, Jerome
American Board of Internal Medicine
jclauser@abim.org

Colvin, Kimberly
University at Albany, SUNY
kcolvin@albany.edu

Cook, Gary
Wisconsin Center for Education Research
hcook@wisc.edu

Cook, H Gary
University of Wisconsin
hcook@wisc.edu

Crabtree, Ashliegh
University of Iowa
Ashleigh-crabtree@uiowa.edu

Croft, Michelle
ACT, Inc.
mcroft@act.org

Cubbellotti, Stephen
American Board of Internal Medicine
scubbellotti@abim.org

Cui, Mengyao
Florida State University
mc09u@my.fsu.edu

Cui, Ying
University of Alberta
yc@ualberta.ca

Curtis, Nicholas
James Madison University
curtisna@dukes.jmu.edu

D

Dadey, Nathan
Center for Assessments
ndadey@nciea.org

Dai, Shenghai
Indiana University Bloomington
dais@indiana.edu

Davidson, Anne
Anne H. Davidson
annehdavidson@gmail.com

Davis, Laurie
Pearson
laurie@davistx.com

Davis-Becker, Susan
ACS Ventures, LLC
sdavisbecker@acsventures.com
sdavisbecker@gmail.com

Davison, Mark
University of Minnesota
mld@umn.edu

de la Torre, Jimmy
The University of Hong Kong
j.delatorre@hku.hk

de Leng, Wendy
Institute of Medical Education Research Rotterdam (iMERR), Erasmus Medical Center
w.deleng@erasmusmc.nl

De Lisle, Jerome
The University of the West Indies
jeromedelisle@yahoo.com

Dell-Ross, Theresa
Georgia State University
tdellross1@gsu.edu

Denbleyker, Johnny
Houghton Mifflin Harcourt
lakeway01@yahoo.com

Deng, Nina
Measured Progress
Deng.Nina@measuredprogress.org

Deters, Lauren
edCount, LLC
ldeters@edcount.com

Diao, Qi
Pacific Metrics Corporation
qdiao@pacificmetrics.com

Dickey, Jacqueline
University of Massachusetts, Amherst
jacqueline.dickey@gmail.com

Dimitrov, Dimiter
National Center for Assessment
ddimitro@gmu.edu

DiTrapani, Jack
The Ohio State University
ditrapani.4@osu.edu

Domaleski, Chris
National Center for the Improvement of Educational Assessment
CDomaleski@nciea.org

Donald, Ellen
Florida Gulf Coast University
ekwill@fgcu.edu

Donoghue, John
Educational Testing Service
jdonoghue@ets.org

Dorans, Neil
Educational Testing Service
ndorans@ets.org

Dudek, Christopher
Rutgers, the State University of New Jersey
cdudek@scarletmail.rutgers.edu

Dunlea, Jamie
British Council
jamie.dunlea@britishcouncil.org

Dunn, Karen
British Council
karen.dunn@britishcouncil.org

Dwyer, Andrew
The American Board of Pediatrics
adwyer@abpeds.org

Dyer, Jarret
College of DuPage
dyerja@cod.edu

E

Eckerly, Carol
Alpine Testing Solutions
caroleckerly@gmail.com

Egan, Karla
EdMetric LLC
karla_egan@edmetricllc.com

Elliott, Stephen
Arizona State University
steve_elliott@asu.edu

Ellis, Rebecca
Michigan State University
ellisre2.msu@gmail.com

Ercikan, Kadriye
University of British Columbia
kadriye.ercikan@ubc.ca

Ezzelle, Carol
National Board for Professional Teaching Standards
ezzelle@gmail.com

F

Fahle, Erin
Stanford University
efahle@stanford.edu

Fan, Fen
National Commission on Certification of Physician Assistants
ffan2010@gmail.com

Fan, Meichu
ACT, Inc.
meichu.fan@act.org

Fang, Yu
ACT, Inc.
yu.fang@act.org

Feinberg, Richard
National Board of Medical Examiners
rfeinberg@nbme.org

Ferrara, Steve
Measured Progress
Ferrara.Steve@measuredprogress.org

Fina, Anthony
Iowa Testing Programs, University of Iowa
anthony-fina@uiowa.edu

Finch, Holmes
Ball State University
whfinch@bsu.edu

Fincher, Melissa
Georgia Department of Education
mfincher@doe.k12.ga.us

Finn, Bridgid
Educational Testing Service
bfinn@ets.org

FitzPatrick, Beverly FitzPatrick
Memorial University
bfitzpatrick@mun.ca

Fitzpatrick, Joseph
University of Kansas
jfitz@ku.edu

Foelber, Kelly
James Madison University
foelbekj@jmu.edu

Forte, Ellen
edCount, LLC
eforte@edcount.com

Fox, Jean-Paul
University of Twente
j.p.fox@utwente.nl

French, Brian
Washington State University
frenchb@wsu.edu

Freund, Rebecca
University of California, Berkeley
rlfreund@berkeley.edu

Frey, Sharon
Houghton Mifflin Harcourt
sharon.frey@hmhco.com

Fu, Yanyan
The University of North Carolina at Greensboro
y_fu2@uncg.edu

Fujimoto, Ken
Loyola University Chicago
kfujimoto@luc.edu

G

Gaj, Shameem
Educational Testing Service
sgaj@ets.org

Geisinger, Kurt
Buros Center of Testing
kgeisinger@buros.org

Gierl, Mark
University of Alberta
mark.gierl@ualberta.ca

Gitomer, Drew
Rutgers University Graduate School of Education
drew.gitomer@gse.rutgers.edu

Glazer, Nancy
Educational Testing Service
nglazer@ets.org

Gong, Brian
Center for Assessment
bgong@nciea.org

Gotch, Chad
Washington State University
cgotch@wsu.edu

Graf, Edith Aurora
Educational Testing Service
agraf@ets.org

Guo, Hongwen
Educational Testing Service
hguo@ets.org

Guo, Qi
University of Alberta
qig@ualberta.ca

Guo, Shaoyang
Jiangxi Normal University
guoshaoyang1992@outlook.com

Guo, Wenjing
City University of New York
wguo@gradcenter.cuny.edu

H

Halpin, Peter
New York University
peter.halpin@nyu.edu

Han, HyunSuk
University of Florida
hh22369@ufl.edu

HAN, YUTING
School of Psychology, Jiangxi Normal University
236181020@qq.com

Han, Zhuangzhuang
Teachers College, Columbia University
zh2198@tc.columbia.edu

Hansen, Mark
University of California, Los Angeles & CRESST.
markhansen@ucla.edu

Hao, Jiangang
ETS
jhao@ets.org

Harik, Polina
National Board of Medical Examiners
pharik@nbme.org

Harrell, Lauren
National Center for Education Statistics
Lauren.Harrell@ed.gov

Hayes, Stacy
Discovery Education
Stacy_Hayes@discovery.com

He, Qiwei
ETS
qhe@ets.org

He, Wei
NWEA
wei.he@nwea.org

He, Yinhong
Beijing Normal University
roheyinhong@163.com

Ho, Andrew
Harvard Graduate School of Education
andrew_ho@gse.harvard.edu

Ho, Tsung-Han
ETS
tho@ets.org

Holden, LaTasha
Princeton
latasha@Princeton.EDU

Hou, Likun
Educational Testing Service
lhou@ets.org

Hu, Bo
The University of Kansas
who.bo@ku.edu

Huh, Nooree
ACT, Inc.
nooree.huh@act.org

Hunter, Charles
Georgia State University
chunter1@gsu.edu

I

Iaconangelo, Charles
Rutgers, The State University of New Jersey
charles.iaconangelo@gmail.com

Ikoma, Sakiko
American Institutes for Research
sikoma@air.org

Insko, William
Houghton Mifflin Harcourt
bill.insko@hmhco.com

Irvin, Phillip
University of Oregon
pirvin@urogeon.edu

Ito, Kyoko
Defense Manpower Data Center
kyoko.ito.civ@mail.mil

J

Jang, Eunice Eunhee
Ontario Institute for Studies in Education, University of Toronto
eun.jang@utoronto.ca

Jang, Yoonsun
University of Georgia
todekfr@gmail.com

Jennings, Jeremy
University of Georgia
jkjennings@gmail.com

Jewsbury, Paul
Educational Testing Service
pjewsbury@ets.org

Jia, Yue (Helena)
ETS
yjia@ets.org

Jiang, Jing
Boston College
jiangjc@bc.edu

Jiang, Yanlin
Educational Testing Service
yjiang@ets.org

Jiang, Yanming
Educational Testing Service
yxjiang@ets.org

Jiang, Zhehan
University of Kansas
zjiang4@ku.edu

Jin, Kuan-Yu
Education University of Hong Kong
kyjin@eduhk.hk

Jin, Rong
Houghton Mifflin Harcourt Publishing
Rong.Jin@hmhco.com

Jin, Ying
Association of American Medical Colleges
yjin@aamc.org

Johnson, Evelyn
Boise State University
evelynjohnson@boisestate.edu

Jonson, Jessica
Buros Center for Testing - University of Nebraska-Lincoln
jjonson@buros.org

Jorion, Natalie
Pearson VUE
talie.jorion@gmail.com

Ju, Unhee
Michigan State University
juunhee@msu.edu

Jun, HeaWon
Georgia Institute of Technology
hjun06@gmail.com

Jung, HyunJoo
University of Massachusetts Amherst
smartcookieno1@gmail.com

K

Kaliski, Pamela
College Board
ppfluger@collegeboard.org

Kamata, Akihito
Southern Methodist University
akamata@smu.edu

Kane, Michael
Educational Testing Service
mkane@ets.org

Kang, Yoonjeong
American Institutes for Research
yoonjeongkang94@gmail.com

Kannan, Priya
Educational Testing Service
pkannan@ets.org

Kao, Shu-chuan
Pearson
shu-chuan.kao@pearson.com

Kaplan, David
University of Wisonsin-Madison
david.kaplan@wisc.edu

Kaplan, Mehmet
The Turkish Ministry of National Education
mehmet.kaplan2@gmail.com

Kapoor, Shalini
ACT
shalinikapoor.ia@gmail.com

Kara, Yusuf
Anadolu University
yusufkara@anadolu.edu.tr

KARADAVUT, TUGBA
UNIVERSITY OF GEORGIA
TUGBA-MAT@HOTMAIL.COM

Karvonen, Meagan
University of Kansas
karvonen@ku.edu

Keller, Lisa
University of Massachusetts Amherst
lkeller@umass.edu

Keuning, Trynke
University of Twente
t.keuning@utwente.nl

Khorramdel, Lale
ETS
lkhorramdel@ets.org

Kim, Do-Hong
University of North Carolina at Charlotte
dkim15@uncc.edu

Kim, Dong-In
Data Recognition Corporation
dkim@datarecognitioncorp.com

Kim, Han Yi
Measured Progress
Kim.HanYi@measuredprogress.org

Kim, Hyung Jin
The University of Iowa
hyungjin-kim@uiowa.edu

Kim, Ja Young
TEPS Center Seoul National University
jaykim319@gmail.com

Kim, JiYoon
Talent Assessment Institute (TAI)
bloomjykim@gmail.com

Kim, Jungnam
NBCE
jungnam95@hotmail.com

Kim, Meereem
University of Georgia
meereemkim@gmail.com

Kim, Min Sung
Buros Center for Testing
mkim@buros.org

Kim, Se-Kang
Fordham University
sekim@fordham.edu

Kim, Seock-Ho
The University of Georgia
shkim@uga.edu

Kim, Seohyun
The University of Georgia
seohyun@uga.edu

Kim, Sooyeon
Educational Testing Service
skim@ets.org

Kim, Stella
The University of Iowa
stella-kim@uiowa.edu

Kim, Young Yee
American Institutes for Research
ykim@air.org

Kim, YoungKoung
The College Board
ykim@collegeboard.org

King, David
Pacific Metrics Corporation
dking@pacificmetrics.com

Kitmitto, Sami
American Institutes for Research
skitmitto@air.org

Kleper, Dvir
National Institute for Testing and Evaluation
dvir@nite.org.il

Konold, Tim
University of Virginia
Konold@Virginia.edu

Kroehne, Ulf
German Institute for International Educational Research (DIPF)
kroehne@dipf.de

Kuhfeld, Megan
University of Texas at Austin
megan.kuhfeld@gmail.com

Kunina-Habenicht, Olga
German Institute for International Educational Research (DIPF), Centre for International Student Assessment (ZIB)
Olga.Kunina-Habenicht@dipf.de

Kuo, Tzu-Chun
Southern Illinois University Carbondale
tckuo@siu.edu

Kwak, Minho
University of Georgia
mk59520@uga.edu

L

LaMar, Michelle
Educational Testing Service
mlamar@ets.org

Lamsal, Sunil
Pearson VUE
sunil.lamsal@pearson.com

Lane, Suzanne
University of Pittsburgh
sl@pitt.edu

Lathrop, Quinn
Advanced Computing and Data Science Lab Pearson
quinn.lathrop@pearson.com

Latifi, Syed
University of Alberta
syed.latifi@ualberta.ca

Lazendic, Goran
Australian Curriculum, Assessment and Reporting Authority (ACARA)
goran.lazendic@acara.edu.au

Lee, HyeSun
California State University Channel Islands
hyesun.kj.lee@gmail.com

Lee, Sung-Hyuck
ACT
sung.lee@act.org

Lehrfeld, Jonathan
Council for Aid to Education
jlehrfeld@cae.org

Lei, Ming
American Institutes for Research
mlei@air.org

Lekwa, Adam
Rutgers, the State University of New Jersey
adam.lekwa@rutgers.edu

Lewis, Daniel
Pacific Metrics Corporation
dlewis5000@sbcglobal.net

Li, Chen
University of Maryland, College Park
lc1210@umd.edu

Li, Jie
McGraw-Hill Education
jie.li@mheducation.com

Li, Min
University of Washington at Seattle
minli@uw.edu

Li, Xiao
University of Illinois at Urbana-Champaign
jnsxlx@gmail.com

Li, Xiaoran
University of Connecticut
xiaoran.li@uconn.edu

Li, Xueming
Northwest Evaluation Association
ppxuemingqq@gmail.com

Li, Zhen
Government of Newfoundland and Labrador
liza0616@hotmail.com

Liang, Longjuan
Educational Testing Service
LYLiang@ets.org

Liao, Chi-wen
Educational Testing Service
cliao@ets.org

Liao, Dandan
University of Maryland
dliao@air.org

Liao, Jue
University of California, Los Angeles
jliao2014@ucla.edu

Liao, Manqian
University of Maryland College Park
mancyliao@gmail.com

Lim, Hwanggyu
University of Massachusetts Amherst
hglim83@gmail.com

Lin, Peng
Educational Testing Service
plin@ets.org

Lin, Songbai
Educational Testing Service
slin@ets.org

Ling, Guangming
Educational Testing Service
gling@ets.org

Ling, Guangming
Educational Testing Service
gling@ets.org

Liu, Cheng
University of Notre Dame
cliu7@nd.edu

Liu, Jinghua
Enrollment Management Association
jliu@enrollment.org

Liu, Junhui
Educational Test Service
junehui.liu@gmail.com

Liu, Lei
ETS
lliu001@ets.org

Liu, Peiyan
University of Denver
peiyan.liu@du.edu

Liu, Ren
University of Florida
liurenking@ufl.edu

Liu, Ruitao
ACT
ruitao.liu@act.org

Liu, Shuangshuang
Educational Testing Service
sliu002@ets.org

Liu, Xiang
Teachers College, Columbia University
xl2438@tc.columbia.edu

LIU, XIANGDONG
THE UNIVERISTY OF IOWA
xiangdong-liu@uiowa.edu

Liu, Xiaoxiao
School of Psychology, Beijing Normal University
psyliuxiaoxiao@126.com

Liu, Xin
Ascend Learning
lucy.xin.liu@gmail.com

Liu, Yang
University of California, Merced
yliu85@ucmerced.edu

Liu, Yuming
Educational Testing Service
yliu@ets.org

Lockwood, J. R.
Educational Testing Service
jrlockwood@ets.org

Lopez, Alexis
Educational Testing Service
alopez@ets.org

Lu, Lucy
Department of Education, New South Wales, Australia
lucy.lu@det.nsw.edu.au

Lu, Ying
Educational Testing Service
ylu@ets.org

Luo, Xiao
National Council of State Boards of Nursing
xluo@ncsbn.org

Luo, Xin
Northwest Evaluation Association (NWEA)
luoxin1@msu.edu
charonluo@gmail.com

Lyons, Susan
National Center for the Improvement of Educational Assessment
slyons@nciea.org

M

Ma, Wenchao
Rutgers, The State University of New Jersey
wenchao.ma@rutgers.edu

Madison, Matthew
University of California - Los Angeles
mjmadison@ucla.edu

Malatesta, Jaime
University of Iowa
jaime-malatesta@uiowa.edu

Mao, Xia
Pearson
xia.mao@pearson.com

Margolis, Melissa
National Board of Medical Examiners
mmargolis@nbme.org

Markle, Ross
Educational Testing Service
rmarkle@ets.org

Mason, James
University of California at BerkEley
jmason888@berkeley.edu

Masters, Jessica
Measured Progress
jessica.masters@researchmattersllc.com

Matlock, Ki
Oklahoma State University
ki.matlock@okstate.edu

Mattern, Krista
ACT, Inc.
krista.mattern@act.org

Maul, Andrew
University of California, Santa Barbara
amaul@education.ucsb.edu

Maynes, Dennis
Caveon, LLC
dennis.maynes@caveon.com

McCaffrey, Daniel
ETS
dmccaffrey@ets.org

McCall, Marty
Smarter Balanced Assessment Consortium
marty.mccall@smarterbalanced.org

McJunkin, Linette
Computerized Assessments & Learning
lmcjunkin@caltesting.org

McMillan, James
Virginia Commonwealth University
jhmcmill@vcu.edu

McNamara, Andrea
Fordham University
andreamcnamara@gmail.com

McNaughton, Tara
Measurement, Inc.
tmcnaughton@measinc.com

Meijer, Rob
University of Groningen
r.r.meijer@rug.nl

Meng, Huijuan
Graduate Management Admission Council® (GMAC®)
hmeng@gmac.com

Michaels, Hillary
HumRRO
hmichaels@humrro.org

Miller, Sherral
College Board
shmiller@collegeboard.org

Minchen, Nathan
Rutgers, The State University of New Jersey
nathan.minchen@gse.rutgers.edu

Monroe, Scott
University of Massachusetts Amherst
smonroe@educ.umass.edu

Montoya, Amanda
The Ohio State University
montoya.29@osu.edu

Morell, Monica
University of Maryland
mmorell@umd.edu

Morris, Carrie
University of Iowa
carrie-morris@uiowa.edu

Moses, Tim
College Board
tmoses@collegeboard.org

N

Nash, Brooke
University of Kansas - Center for Educational Testing and Evaluation
bnash@ku.edu

Naumann, Johannes
Goethe University
j.naumann@em.uni-frankfurt.de

Nebelsick-Gullett, Lori
edCount, LLC
lnebelsick-gullett@edCount.com

Nese, Joseph
University of Oregon
jnese@uoregon.edu

Nichols, Paul
ACT
paul.nichols@act.org

Nie, Xugang
Beijing Normal University
niexugang@163.com

Niessen, Susan
University of Groningen
a.s.m.niessen@rug.nl

NIu, Luping
The University of Texas at Austin
newl787@gmail.com

Northcutt, Curtis
MIT
cgn@mit.edu

O

Ogut, Burhan
American Institutes for Research
bogut@air.org

Oh, Hyeonjoo
ETS
hoh@ets.org

O'Leary, Timothy
Melbourne Graduate School of Education
olearyt@student.unimelb.edu.au

Oliveri, Maria
Educational Testing Service
moliveri@ets.org

Olsen, Joseph
Brigham Young University
joseph_olsen@byu.edu

Olson, Evan
University of Maryland
eolson01@yahoo.com

Oranje, Andreas
ETS
aoranje@ets.org

O'Reilly, Tenaha
Educational Testing Service
toreilly@ets.org

Ou, Lu
Penn State
lzo114@psu.edu

P

Pace, Jesse
University of Kansas
Jesse.Pace@ku.edu

Padellaro, Frank
University of Massachusetts
fpadellaro@umass.edu

Paek, Pamela
ACT
Pamela.Paek@act.org

Pak, Seohong
The University of Iowa
seohong-pak@uiowa.edu

Park, Bitnara
American Institutes for Research
bpark@air.org

Park, Jiyoon
Federation of State Boards of Physical Therapy
jpark@fsbpt.org

Park, Ryoungsun
Wayne State University
fy3504@wayne.edu

Patarapichayatham, Chalie
Southern Methodist University
chalie.pt@gmail.com

Patelis, Thanos
Center for Assessment
tpatelis@nciea.org

Patsula, Liane
Medical Council of Canada
Lpatsula@mcc.ca

Perie, Marianne
University of Kansas
Center for Educational Testing and Evaluation
mperie@ku.edu

Petway, Kevin
Educational Testing Service
kpetway@ets.org

Phadke, Chaitali
University of Minnesota
phadk011@umn.edu

Pham, Duy
University of Massachussets Amherst
dpham@umass.edu

Phelps, Geoffrey
Educational Testing Service (ETS)
gphelps@ets.org

Por, Han Hui
Educational Testing Service
hpor@ets.org

Q

Qian, Jiahe
Educational Testing Service
jqian58@gmail.com

Qian, Xiaoyu
Educational Testing Service
xqian@ets.org

QIU, Xue-Lan
The Education University of Hong Kong
xlqiu@eduhk.hk

QIU, YUXI
UNIVERSITY OF FLORIDA
yqiu2013@ufl.edu

Quinn, David
University of Southern California
quinnd@usc.edu

R

Rabinowitz, Stanley
Australian Curriculum, Assessment and Reporting Authority (ACARA)
stanley.rabinowitz@acara.edu.au

Rafferty, Anna
Carleton College
arafferty@carleton.edu

Ramanarayanan, Vikram
Educational Testing Service
VRAMANARAYANAN@ets.org

Rankin, Angelica
Project Lead the Way
rankinangelicad@gmail.com

Rawls, Anita
College Board
arawls@collegeboard.org

Raymond, Mark
National Board of Medical Examiners
MRaymond@nbme.org

Reid, Jerry
The American Registry of Radiologic Technologists
jerry.reid@arrt.org

Ren, Hao
Pacific Metrics Corporation
hren@pacificmetrics.com

Reshetnyak, Evgeniya
Fordham University
ereshetnyak@fordham.edu

Rick, Francis
University of Massachusetts, Amherst
frick@umass.edu

Rickels, Heather
University of Iowa, Iowa Testing Programs
heather-rickels@uiowa.edu

Rijmen, Frank
American Institutes for Research
frijmen@air.org

Roberts, William
National Board of Osteopathic Medical Examiners
broberts@nbome.org

Rodriguez, Michael
University of Minnesota
mcrdz@umn.edu

Roduta Roberts, Mary
University of Alberta
mroberts@ualberta.ca

Rollins III, Jonathan
The University of North Carolina at Greensboro
jdrollin@uncg.edu

Rome, Logan
University of Wisconsin-Milwaukee
larome@uwm.edu

Roohr, Katrina
Educational Testing Service
kroohr@ets.org

Rosen, Yigal
Harvard University
yigal_rosen@harvard.edu

Runyon, Christopher
The University of Texas at Austin
runyon.christopher@utexas.edu

Rupp, Andre
Educational Testing Service
arupp@ets.org

S

Sabatini, John
Educational Testing Service
jsabatini@ets.org

Sahin, Fusun
University at Albany, State University of New York
fsahin@albany.edu

Sales, Adam
University of Texas College of Education
asales@utexas.edu

S | Contact Information for Individual and Coordinated Sessions First Authors

Santos, Kevin
University of the Philippines- Diliman
kpsantos1@up.edu.ph

Santos, Kevin Carl
University of the Philippines-Diliman
kpsantos1@up.edu.ph

Sato, Edynn
Sato Education Consulting LLC
edynn_s@yahoo.com

Schneider, Christina
NWEA
christina.schneider@nwea.org

Schultz, Matthew
American Institute of Certified Public Accountants
maschultz@aicpa.org

Schwartz, Robert
Pearson
schwartz.bob@gmail.com

Seaton, Daniel
Harvard Universtiy
daniel_seaton@harvard.edu

Seburn, Mary
Quantiful, LLC.
mseburn@quantiful.com

Seltzer, Michael
UCLA
mseltzer@ucla.edu

Sessoms, John
University of North Carolina at Greensboro
jcsessom@uncg.edu

Setzer, Carl
AICPA
csetzer@aicpa.org

Sgammato, Adrienne
Educational Testing Service
asgammato@ets.org

Shao, Can
National Board of Osteopathic Medical Examiners
cshao@nd.edu

Sharp, Amy
University of Washington
sharpa2@myuw.net

Shear, Benjamin
University of Colorado Boulder
benjamin.shear@colorado.edu

Sheehan, Kathleen
Educational Testing Service
ksheehan@ets.org

Shi, Dingjing
University of Virginia
ds4ue@virginia.edu

Shin, Nami
CRESST/UCLA
shin@cresst.org

Sinharay, Sandip
ETS
ssinharay@ets.org

Sireci, Stephen
University of Massachusetts-Amherst
sireci@acad.umass.edu

Skar, Gustaf Bernhard Uno
Norwegian University of Science and Technology
gustaf.b.skar@ntnu.no

Slater, Steve
Oregon Department of Education
steve.slater@state.or.us

Smith, Jessalyn
DRC
smith.jessalyn@gmail.com

Smith, Weldon
University of Nebraska-Lincoln
weldon@huskers.unl.edu

Song, Yi
Educational Testing Service
ysong@ets.org

Sorrel, Miguel
Universidad Autónoma de Madrid
miguel.sorrel@uam.es

Stafford, Rose
The University of Texas at Austin
rose.stafford@utexas.edu

Steedle, Jeffrey
Pearson
jtsteedle@gmail.com

Stegers-Jager, Karen
Institute of Medical Education Research Rotterdam (iMERR), Erasmus Medical Center
k.stegers-jager@erasmusmc.nl

Steinberg, Jonathan
Educational Testing Service
jsteinberg@ets.org

Stevens, Joseph
University of Oregon
stevensj@uoregon.edu

Suh, Hongwook
ACT
hongwooks@gmail.com

Sullivan, Meghan
University of Kansas
meg.sullivan@ku.edu

Susadya, Laurentius
University of Iowa
laurentius-susadya@uiowa.edu

Sweeney, Kevin
The College Board
ksweeney@collegeboard.org

T

Tan, Xuan (Adele)
Educational Testing Service
atan@ets.org

Tang, Wei
University of Alberta
bulut@ualberta.ca

Tang, Xiaodan
University of Illinois at Chicago
xtang23@uic.edu

Tannenbaum, Richard
Educational Testing Service (ETS)
rtannenbaum@ets.org

Tao, Shuqin
Curriculum Associates
shuqin.tao@gmail.com

Tao, Wei
ACT, Inc.
wei.tao@act.org

Templin, Jonathan
University of Kansas
jtemplin@ku.edu

Thomas, Jay
ACT, Inc.
Jay.Thomas@act.org

Thomas Pitts, Robyn
University of North Carolina at Greensboro
rlthoma2@uncg.edu

Timperley, Helen
The University of Auckland
h.timperley@auckland.ac.nz

Tjoe, Hartono
Penn State Berks
hartonotjoe@hotmail.com

Tokac, Umit
Florida State University
ut08@my.fsu.edu

Topczewski, Anna
GED Testing Service
anna.topczewski@gedtestingservice.com

Torres Irribarra, David
MIDE, Pontificia Universidad Católica de Chile
davidtorres@uc.cl

Traynor, Anne
Purdue University
atraynor@purdue.edu

Tzou, Hueying
National University of Tainan
tzou@mail.nutn.edu.tw

V

van der Linden, Wim
Pacific Metrics Corporation
wvanderlinden@pacificmetrics.com

van Geel, Marieke
University of Twente
m.j.m.vangeel@utwente.nl

Vasquez-Colina, Maria
Florida Atlantic University
mvasque3@fau.edu

Vispoel, Walter
University of Iowa
walter-vispoel@uiowa.edu

von Davier, Alina
ETS
avondavier@ets.org

von Davier, Matthias
ETS
mvondavier@ets.org

W

Wallin, Gabriel
Umeå School of Business and Economics, Umeå University
gabriel.wallin@umu.se

Wan, Lei
College Board
wanlei77@yahoo.com

wang, Aijun
federation of state boards of physical therapy
awang@fsbpt.org

Wang, Chunxin
ACT Inc.
ann.wang@act.org

Wang, Keyin
Michigan State University
keyinw0323@gmail.com

Wang, Lin
Educational Testing Service
lwang@ets.org

Wang, Qinjun
University of Minnesota
wang4314@umn.edu

Wang, Shichao
The University of Iowa
shichao-wang@uiowa.edu

Wang, Shiyu
University of Georgia
swang44@uga.edu

Wang, Shudong
NWEA
shudong.wang@NWEA.org

Wang, Ting
ETS
twang001@ets.org

Wang, Xi
Measured Progress
smilingwx2010@gmail.com

Wang, Xiaolin
Indiana University - Bloomington
xw41@indiana.edu

Wang, Zhen
Educational Testing Service
Jwang@ETS.ORG

Wang, Zhuoran
University of Minnesota, Twin Cities
wang5105@umn.edu

Wang, Zuowei
Educational Testing Service
zwang@ets.org

Way, Denny
Pearson
Denny.way@pearson.com

Weeks, Jonathan
ETS
jweeks@ets.org

Welch, Catherine
University of Iowa
catherine-welch@uiowa.edu

Wells, Craig
University of Massachusetts Amherst
cswells@educ.umass.edu

Wendler, Cathy
Educational Testing Service
cwendler@ets.org

West, Martin
Harvard Graduate School of Education
martin_west@gse.harvard.edu

Wiberg, Marie
Umeå University
marie.wiberg@umu.se

Widiatmo, Heru
ACT, Inc.
heru.widiatmo@act.org

Wiley, Andrew
ACS Ventures
awiley@acsventures.com

Wilke, Ryan
Florida State University
ryanawilke@gmail.com

Williamson, Gary
MetaMetrics
gwilliamson@lexile.com

Wilson, Joshua
University of Delaware
joshwils@udel.edu

Wind, Stefanie
The University of Alabama
swind@ua.edu

Wise, Steven
Northwest Evaluation Association
steve.wise@nwea.org

Wolf, Raffaela
Pearson Vue
raffaela.wolf@gmail.com

Wood, Scott
Pacific Metrics Corporation
swood@pacificmetrics.com

Wu, Brad
Pearson VUE
brad.wu@pearson.com

Wu, Meng
ETS
mwu@ets.org

Wu, Yi-Fang
ACT, Inc.
Yi-Fang.Wu@act.org

Wyse, Adam
The American Registry of Radiologic Technologists
adam.wyse@arrt.org

X

Xi, Nuo
Educational Testing Service
nxi@ets.org

Xie, Chao
American Institutes for Research
cxie@air.org

Xie, Qing
University of Iowa/ ACT, Inc
qing-xie@uiowa.edu

Xin, Tao
Beijing Normal University
xintao@bnu.edu.cn

Xing, Kuan
University of Illinois at Chicago
kuanxing83@gmail.com

Xiong, Xinhui
American Institute of Certified Public Accountants
xxiong@aicpa.org

Xu, Menglin
The Ohio State University
xumenglin920@gmail.com

Xu, Wei
University of Florida
x.wei1007@gmail.com

Xu, Xueli
ETS
xxu@ets.org

Y

Yakar, Levent
Hacettepe University
l_yakar@hotmail.com

Yao, Lihua
Defense ManPower Data Center
Lihua.Yao.civ@mail.mil

Yao, Lili
Educational Testing Service
lyao@ets.org

Ye, Sangbeak
University of Illinois - Urbana Champaign
sye3@illinois.edu

Yi, Qing
ACT
qing.yi@act.org

Yigit, Hulya Duygu
Rutgers, The State University of New Jersey
hy261@scarletmail.rutgers.edu

Yoo, Hanwook
Educational Testing Service
hyoo@ets.org

Yu, Lei
Measured Progress
yu.lei@measuredprogress.org

Z

Zhan, Peida
Beijing Normal University
pdzhan@gmail.com

Zhang, Liru
Delaware Department of Education
liru.zhang@doe.k12.de.us

Zhang, Mingcai
Michigan State University
zhangmc@msu.edu

Zhang, Mo
ETS
mzhang@ets.org

Zhang, Susu
University of Illinois at Urbana-Champaign
szhan105@illinois.edu

Zhang, Xinxin
University of Alberta
xinxin4@ualberta.ca

Zhang, Xue
Northeast Normal University
zhangx815@nenu.edu.cn

Zhao, Xinchu
University of South Carolina
xinchu@email.sc.edu

Zheng, Guoguo
University of Georgia
ggzheng@uga.edu

Zheng, Yating
University of Maryland at College Park
yzheng12@umd.edu

Zisk, Robert
Rutgers University Graduate School of Education
robert.zisk@gse.rutgers.edu

Zlatkin-Troitschanskaia, Olga
Johannes Gutenberg-University Mainz
brueckner@uni-mainz.de

Zopluoglu, Cengiz
University of Miami
c.zopluoglu@miami.edu

NCME 2017 • Schedule-At-A-Glance

Time	Room	Type	ID	Title
Wednesday, April 26, 2017				
8:00 AM–12:00 PM	Salon K	TS	AA	Vertical scaling methodologies, applications and research
8:00 AM–12:00 PM	Salon L	TS	BB	An introduction to linking and equating in R
8:00 AM–12:00 PM	Salon M	TS	CC	Rubrics for classroom asssessment: perils of practice and how to avoid them
8:00 AM–5:00 PM	Salon A	TS	DD	Bayesian networks in educational assessment
8:00 AM–5:00 PM	Salon B	TS	EE	Shadow-test approach to adaptive testing
8:00 AM–5:00 PM	Salon C	TS	FF	Cognitive diagnostic modeling: A General framework approach and its implementation in R
8:00 AM–5:00 PM	Salon D	TS	GG	Conceptual frameworks for aligning items to ALDs to enhance validity arguments
1:00 PM–5:00 PM	Salon K	TS	HH	The history of educational measurement in America: Origins to 1950
1:00 PM–5:00 PM	Salon L	TS	II	Landing your dream job for graduate students
1:00 PM–5:00 PM	Salon M	TS	JJ	Data rich, information poor: Navigating data use in a balanced assessment system
1:00 PM–5:00 PM	Conference Room 12	TS	KK	An introduction to R for quantitative methods
1:00 PM–5:00 PM	Conference Room 1&2	TS	LL	Analyzing NAEP data using plausible values and marginal estimation with AM
Thursday, April 27, 2017				
8:00 AM–12:00 PM	Salon L	TS	MM	Evidenced-centered design and computational psychometrics solution for game/simulation-based assessments
8:00 AM–12:00 PM	Salon M	TS	NN	Moving from paper to online assessment: Psychometric, content, and classroom considerations
8:00 AM–5:00 PM	Salon A	TS	OO	Diagnostic classification models: Theory, methods, and applications
8:00 AM–5:00 PM	Salon B	TS	PP	Interpersonal and intrapersonal skills assessment: Design, development, scoring, and reporting
8:00 AM–5:00 PM	Salon C	TS	QQ	Bayesian estimation of item response theory model parameters using OpenBUGS and Stan
8:00 AM–5:00 PM	Salon D	TS	RR	An introduction to hierarchical rater models for the analysis of ratings
8:00 AM–5:00 PM	Salon K	TS	SS	A framework and platform for the development of assessment literacy

CS=Coordinated Session • EB= Electronic Board Session
IS= Invited Session • PS= Paper Session • TS=Training Session

Time	Room	Type	ID	Title
1:00 PM–5:00 PM	Salon L	TS	TT	Computerized multistage adaptive testing: Theory and applications
1:00 PM–5:00 PM	Salon M	TS	UU	Using visual displays to inform assessment development and validation
1:00 PM–5:00 PM	Conference Room 12	TS	VV	Evaluating alignment of computer adaptive assessments
1:00 PM–5:00 PM	Conference Room 1&2	TS	WW	A visual introduction to computerized adaptive testing
4:00 PM–7:00 PM	Conference Room 9			NCME Board of Directors Meeting
Friday, April 28, 2017				
6:30 AM–7:30 AM	Rio Vista Room			Yoga
8:15 AM–10:15 AM	Salon A	IS	A1	The Ocean of Data from Classroom EdTech: Are Psychometricians Ready?
8:15 AM–10:15 AM	Conference Room 1&2	CS	A2	Challenges in Automated Scoring beyond Features and Models
8:15 AM–10:15 AM	Conference Room 3&4	CS	A3	Cognitively Diagnostic Assessment of Middle School Proportional Reasoning: Development and Analysis
8:15 AM–10:15 AM	Salon B	CS	A4	Modeling Learning Progressions and Informing Practice
8:15 AM–10:15 AM	Salon C	CS	A5	Large-scale social-emotional skills assessment in K-12
8:15 AM–10:15 AM	Salon D	PS	A6	Test Security and Model Fit
8:15 AM–10:15 AM	Salon K	PS	A7	Linking & Equating with IRT
8:15 AM–10:15 AM	Salon L	PS	A8	Validity Issues and International Assessment
8:15 AM–10:15 AM	Salon M	PS	A9	Reliability Issues in IRT
10:35 AM–12:05 PM	Salon EF	IS	B1	Classroom Assessment: Promises, Perils, and Next Steps for Moving Forward
12:25 PM–1:55 PM	Salon A	IS	C1	NCME Award Winners
12:25 PM–1:55 PM	Conference Room 1&2	CS	C2	Bayesian Developments in Modeling and Prediction for LSA Data with Applications
12:25 PM–1:55 PM	Conference Room 3&4	CS	C3	Psychometric Issues In Multidimensional IRT
12:25 PM–1:55 PM	Salon B	CS	C4	Engineered Cut Scores: Aligning Standard Setting Methodology with Contemporary Assessment Design Principles
12:25 PM–1:55 PM	Salon C	CS	C5	Building validity arguments for domestic and international college learning outcomes assessments
12:25 PM–1:55 PM	Salon D	PS	C6	Issues in Model-data (mis)fit

CS=Coordinated Session • EB= Electronic Board Session
IS= Invited Session • PS= Paper Session • TS=Training Session

Time	Room	Type	ID	Title
12:25 PM–1:55 PM	Salon K	PS	C7	Modeling non-cognitive traits with IRT
12:25 PM–1:55 PM	Salon L	PS	C8	Bayesian Modeling
12:25 PM–1:55 PM	Salon M	PS	C9	Automatic Item Generation
12:25 PM–1:55 PM	Salon J	EB	C10	Electronic Board Session 1
2:15 PM–3:45 PM	Salon A	CS	D1	Contemporary Issues with Interim Assessments
2:15 PM–3:45 PM	Conference Room 1&2	CS	D2	Classroom realities that need to be understood to make good assessments
2:15 PM–3:45 PM	Conference Room 3&4	CS	D3	Multilevel Latent Variable Modeling Strategies for Handling Error in Predictors
2:15 PM–3:45 PM	Salon B	CS	D4	Improving Technical Quality in K-12 Computerized-Adaptive Tests
2:15 PM–3:45 PM	Salon C	CS	D5	Reworking the Multiple-choice Item to Improve, Teaching, Learning and Program Evaluation
2:15 PM–3:45 PM	Salon D	PS	D6	Dealing with Missing Data
2:15 PM–3:45 PM	Salon K	PS	D7	Issues in Standard Setting 1
2:15 PM–3:45 PM	Salon L	PS	D8	Validity Issues in Test Design
2:15 PM–3:45 PM	Salon M	PS	D9	Generalized Linear Mixed Models
2:15 PM–3:45 PM	Salon J	EB	D10	GSIC Graduate Student Poster Session 1
4:05 PM–6:05 PM	Salon A	CS	E1	Can we estimate all US school test-score distributions from ordinal proficiency data?
4:05 PM–6:05 PM	Conference Room 1&2	CS	E2	Issues in Human Rater Calibration: How Much is Enough?
4:05 PM–6:05 PM	Conference Room 3&4	CS	E3	Looking Back and Moving Forward on Score Reporting Research and Practice
4:05 PM–6:05 PM	Salon B	CS	E4	New Perspectives on Performing Job Analysis
4:05 PM–6:05 PM	Salon C	CS	E5	Walking a tightrope: Navigating the balance of policy and psychometrics
4:05 PM–6:05 PM	Salon D	PS	E6	DIF session 1
4:05 PM–6:05 PM	Salon K	PS	E7	Test Design issues with Linking
4:05 PM–6:05 PM	Salon L	PS	E8	Test Design isssues with Diagnostic Classification Models
4:05 PM–6:05 PM	Salon M	PS	E9	Applications of Multilevel Modeling
4:05 PM–6:05 PM	Salon J	EB	E10	Electronic Board Session 2
4:30 PM–6:30 PM	Yard House (River Walk) 849 E. Commerce St. San Antonio, TX			Graduate Student Social
6:30 PM–8:00 PM	Salon I			NCME and Division D Reception

Time	Room	Type	ID	Title
Saturday, April 29, 2017				
8:00 AM–10:00 AM	Salon EF			NCME Breakfast and Business Meeting
10:35 AM–12:05 PM	Salon A	IS	F1	Career Award
10:35 AM–12:05 PM	Conference Room 1&2	CS	F2	Student Learning Objectives and the Challenge of Campbell's Law
10:35 AM–12:05 PM	Conference Room 3&4	CS	F3	Mixture and Bayesian IRT Approaches to Modeling Guessing and Beyond
10:35 AM–12:05 PM	Salon B	CS	F4	Perspectives on Quality in Assessment Practice
10:35 AM–12:05 PM	Salon C	CS	F5	Stories Told by Test Cheaters: What we can learn from them
10:35 AM–12:05 PM	Salon D	PS	F6	Test design issues in classroom assessment
10:35 AM–12:05 PM	Salon K	PS	F7	Multidimensional estimation of subscores
10:35 AM–12:05 PM	Salon L	PS	F8	Issues in Standard Setting 2
10:35 AM–12:05 PM	Salon M	PS	F9	Score Reporting
10:35 AM–12:05 PM	Salon J	EB	F10	Electronic Board Session 3
2:45 PM–4:15 PM	Salon A	IS	G1	Peer Review Under the Every Student Succeeds Act of 2015
2:45 PM–4:15 PM	Conference Room 1&2	CS	G2	Connecting Psychometrics and Classroom Assessment to Improve Theory and Practice
2:45 PM–4:15 PM	Conference Room 3&4	CS	G3	Methodological Advances in PISA
2:45 PM–4:15 PM	Salon B	CS	G4	Designing and modeling performance-based assessments of student collaboration
2:45 PM–4:15 PM	Salon C	CS	G5	Issues with Interpreting Measurement Results in Social Sciences: Causality, Dimensionality and Scales
2:45 PM–4:15 PM	Salon D	PS	G6	Models for Polytomous Data
2:45 PM–4:15 PM	Salon K	PS	G7	Generalizability Theory
2:45 PM–4:15 PM	Salon L	PS	G8	Technical Issues with Linking & Equating
2:45 PM–4:15 PM	Salon M	PS	G9	Testing Special Populations
4:35 PM–6:05 PM	Salon A	IS	H1	Psychological and Social Measurement: The Career and Contributions of Benjamin D. Wright
4:35 PM–6:05 PM	Conference Room 1&2	CS	H2	Psychometrics for noncognitive assessment: Unique challenges and some solutions

CS=Coordinated Session • EB= Electronic Board Session
IS= Invited Session • PS= Paper Session • TS=Training Session

Time	Room	Type	ID	Title
4:35 PM–6:05 PM	Conference Room 3&4	CS	H3	Innovative Approaches to Fairly Designing and Developing Noncognitive Measures for Diverse Populations
4:35 PM–6:05 PM	Salon B	CS	H4	Flexible K-12 Assessments Afforded by ESSA: Psychometric Possibilities and Case Studies
4:35 PM–6:05 PM	Salon C	CS	H5	Development and Implementation of a Comprehensive Alignment Evaluation Framework
4:35 PM–6:05 PM	Salon D	PS	H6	DIF Session 2
4:35 PM–6:05 PM	Salon K	PS	H7	Issues with Computerized Assessments
4:35 PM–6:05 PM	Salon L	PS	H8	Reliability and Validity of Classroom Assessment
4:35 PM–6:05 PM	Salon M	PS	H9	Advances in Multilevel Modeling
6:30 PM–8:00 PM	Rio Vista Room			NCME President's Reception
Sunday, April 30, 2017				
8:15 AM–10:15 AM	Salon A	IS	I1	Assessing Student Learning Outcomes in Higher Education
8:15 AM–10:15 AM	Conference Room 1&2	CS	I2	New Development and Applications of Automated Scoring in Educational Assessments
8:15 AM–10:15 AM	Conference Room 3&4	CS	I3	Applications of Regression Concepts to Test Linking
8:15 AM–10:15 AM	Salon B	CS	I4	Issues and Challenges in the Measurement of Noncognitive Skills in Applied Settings
8:15 AM–10:15 AM	Salon C	CS	I5	Putting the Content Back into Validity Evidence
8:15 AM–10:15 AM	Salon D	PS	I6	Test Security Issues
8:15 AM–10:15 AM	Salon K	PS	I7	Multidimensional IRT
8:15 AM–10:15 AM	Salon L	PS	I8	Bayesian Approaches to IRT
8:15 AM–10:15 AM	Salon M	PS	I9	Diagnostic Classification Models
10:35 AM–12:05 PM	Salon A	IS	J1	Fairness in Educational Assessment and Measurement
10:35 AM–12:05 PM	Conference Room 1&2	IS	J2	Measuring Creativity from Classrooms to Large Scale Assessments: Views from Practice to Research and Development of Assessments
10:35 AM–12:05 PM	Conference Room 3&4	CS	J3	Challenges and Opportunities in the Application of Diagnostic Measurement
10:35 AM–12:05 PM	Salon B	CS	J4	Understanding Student Decision Making using Markov Decision Processes
10:35 AM–12:05 PM	Salon C	CS	J5	Understanding the difference between professional and legal expectations: A practitioner's challenge
10:35 AM–12:05 PM	Salon D	PS	J6	Vertical Scaling

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	Salon K	PS	J7	Testlet-based Tests
10:35 AM–12:05 PM	Salon L	PS	J8	Subscore Reliability
10:35 AM–12:05 PM	Salon M	PS	J9	Applications of Structural Equation Models
10:35 AM–12:05 PM	Salon J	EB	J10	Electronic Board Session 4
12:00 PM–2:00 PM	Conference Room 9			NCME Past President's Luncheon
12:25 PM–1:55 PM	Salon A	CS	K1	Multi-Method Validation Using an Integrated Evidence-Centered Design Process
12:25 PM–1:55 PM	Conference Room 1&2	CS	K2	Bayesian Sequential Methods for Adaptive Testing
12:25 PM–1:55 PM	Conference Room 3&4	CS	K3	New measures of digital access, familiarity, efficacy, and engagement for large-scale assessments
12:25 PM–1:55 PM	Salon B	CS	K4	Current Developments in Selective Admission to Higher Education in Europe
12:25 PM–1:55 PM	Salon C	CS	K5	The Impact of Background Knowledge in Reading for Understanding
12:25 PM–1:55 PM	Salon D	PS	K6	Teacher Evaluation
12:25 PM–1:55 PM	Salon K	PS	K7	Linking & Equating with Small Sample Sizes
12:25 PM–1:55 PM	Salon L	PS	K8	Technical Advances in IRT
12:25 PM–1:55 PM	Salon M	PS	K9	Reliability Issues
12:25 PM–1:55 PM	Salon J	EB	K10	Electronic Board Session 5
2:15 PM–3:45 PM	Salon A	CS	L1	Opportunity to Learn: Impact on Large-Scale and Classroom Assessment Design and Interpretation
2:15 PM–3:45 PM	Conference Room 1&2	CS	L2	Understanding, Validating, and Evaluating Assessment: Promoting Sound Decisions through Assessment Literacy
2:15 PM–3:45 PM	Conference Room 3&4	CS	L3	ELPA21: Standard Setting and Score Interpretation
2:15 PM–3:45 PM	Salon B	CS	L4	Not-your-average mode study: NAEP goes tablet
2:15 PM–3:45 PM	Salon C	CS	L5	The alignment onion: Multiple layers to validity considerations
2:15 PM–3:45 PM	Salon D	PS	L6	Structural Equating Modeling
2:15 PM–3:45 PM	Salon K	PS	L7	Computerized Adaptive Testing
2:15 PM–3:45 PM	Salon L	PS	L8	Evaluating Model-data Fit in IRT
2:15 PM–3:45 PM	Salon M	PS	L9	Automated Scoring
2:15 PM–3:45 PM	Salon J	EB	L10	GSIC Graduate Student Poster Session 2

Time	Room	Type	ID	Title
4:00 PM–7:00 PM	Conference Room 10			NCME Board of Directors Meeting
4:05 PM–6:05 PM	Salon A	IS	M1	Diversity Issues in Testing Committee Sponsored Symposium
4:05 PM–6:05 PM	Conference Room 1&2	CS	M2	Assessment in MOOCs: Current and Next Generation Research and Development
4:05 PM–6:05 PM	Conference Room 3&4	CS	M3	Designing and Evaluating Score Reports for Specific Audiences
4:05 PM–6:05 PM	Salon B	CS	M4	Device effects research in NAPLAN transition to online testing
4:05 PM–6:05 PM	Salon C	CS	M5	Testing, testing: What is the fairest score when applicants retake admissions tests?
4:05 PM–6:05 PM	Salon D	PS	M6	DIF Session 3
4:05 PM–6:05 PM	Salon K	PS	M7	Model fit issues with Diagnotic Classification Models
4:05 PM–6:05 PM	Salon L	PS	M8	Applications of Validity Methods
4:05 PM–6:05 PM	Salon M	PS	M9	Reliability of Ratings