NCME national council on measurement in education

National Council on Measurement in Education

2015 Training Sessions
April 15-16

2015 Annual Meeting
April 17-19

InterContinental Hotel
Chicago, Illinois

Table of Contents

NCME Officers

President	Lauress Wise *HumRRO, Seaside, CA*
Vice President	Richard J. Patz *ACT, Iowa City, IA*
Past President	Wim van der Linden *CTB/McGraw-Hill, Monterey, CA*
Executive Officer	Susan Rees *NCME Interim Executive Director, Madison, WI*

NCME Directors

Susan Brookhart
Brookhart Enterprises, LLC, Helena, MT

Amy Hendrickson
The College Board, Newtown, PA

Kristen Huff
Regents Research Fund, Brooklyn, NY

Jennifer L. Kobrin, Secretary
Pearson, Wayne, NJ

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

Cindy Walker
University of Wisconsin-Milwaukee, Milwaukee, WI

Huafang Zhao
Montgomery County Public Schools, Rockville, MD

Editors

Journal of Educational Measurement	Jimmy de la Torre *Rutgers, The State University of New Jersey, New Brunswick, NJ*
Educational Measurement Issues and Practice	Derek Briggs *University of Colorado, Boulder, CO*
NCME Newsletter	Heather M. Buzick *Educational Testing Service, Princeton, NJ*
Website Management Committee	Brett Foley *Alpine Testing Solutions, Denton, NE*

2015 Annual Meeting Chairs

Annual Meeting Program Chairs	Jennifer Randall *University of Massachusetts, Amherst, MA* Ye Tong *Pearson*
Training and Development Committee Chair	Caroline Wiley *HumRRO, Alexandria, VA*
Fitness Run/Walk Directors	Brian F. French *Washington State University, Pullman, WA* Jill van den Heuvel *Alpine Testing Solutions, Hatfield, PA*

NCME Information Desk

The NCME Information Desk is located in the Valencia Foyer at the InterContinental Hotel. Stop by to pick up a ribbon and obtain your bib number for the fun run and walk. It will be open at the following times:

Wednesday, April 15	7:30 AM-4:30 PM
Thursday, April 16	7:30 AM-4:30 PM
Friday, April 17	8:00 AM-4:30 PM
Saturday, April 18	10:00 AM-4:30 PM
Sunday, April 19	8:00 AM-1:00 PM

Proposal Reviewers

Alvaro Arce-Ferrer
Karen Barton
Laine Brandshaw
Kirk Becker*
Paul De Boeck*
Chad Buckendahl*
Gregory Camilli
Allan Cohen
Jerome Dagostino
Alina Davier*
Susan Davis-Becker
Jimmy de la Torre
Jennifer Dunn
Steve Ferrara
Holmes Finch
Ellen Forte
Ronald K. Hambleton
Swaminathan Hariharan
Samuel Haring
Deborah Harris
Robert Henson
Andrew Ho*
Kris Kaase
Lisa Keller
Rob Kirkpatrick
Hollis Lai
Won-chan Lee*
Jacqueline Leighton
Feiming Li
Susan Lottridge
Ric Luecht*
Krista Mattern*
Jennifer Merriman
Patrick Meyer
Kimberly O'Malley
Alan Nicewander
James Olsen
Andreas Oranje
Jose-Luis Padilla
Thanos Patelis
Rich Patz
Mary Pitoniak*
Barbara Plake
Jon Poggio
Jane Rogers
Robert Schwartz
Matthew Schultz
Sandip Sinharay
Stephen Sireci*
William Skorupski
Amanda Soto
Tia Sukin
Tony Thompson
Anna Topczewski
Wim van der Linden
Cindy Walker
Michael Walker*
Changjiang Wang*
Walter (Denny) Way
Jonathan Weeks
Craig Wells*
Cathy Wendler
John Willse
Drew Wiley*
Lauress Wise
Steve Wise
April Zenisky*

*Indicates Expert Panel Chairperson

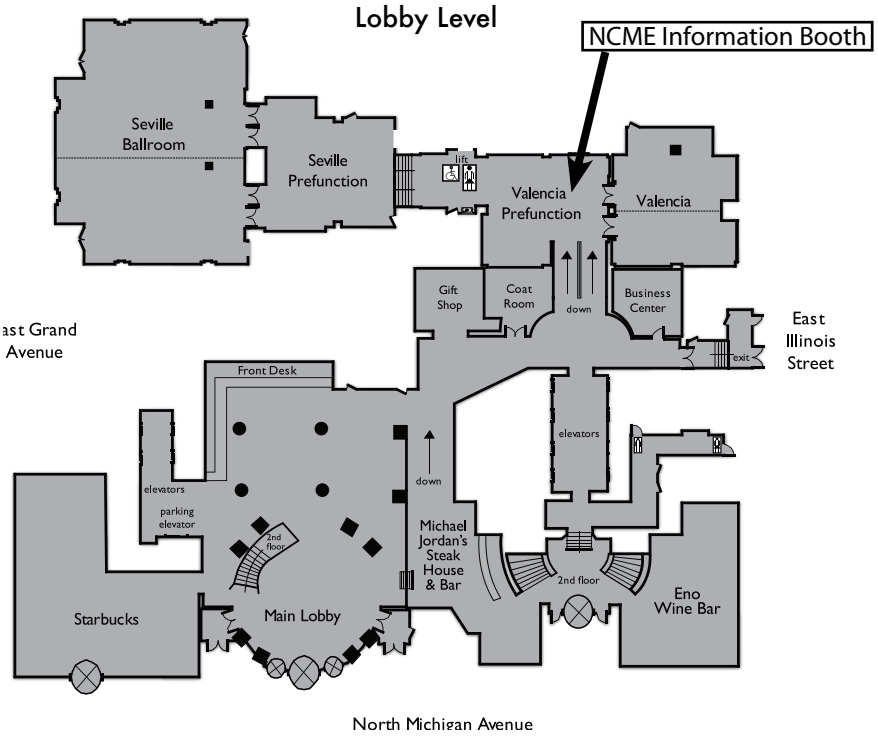
Graduate Student Abstract Reviewers

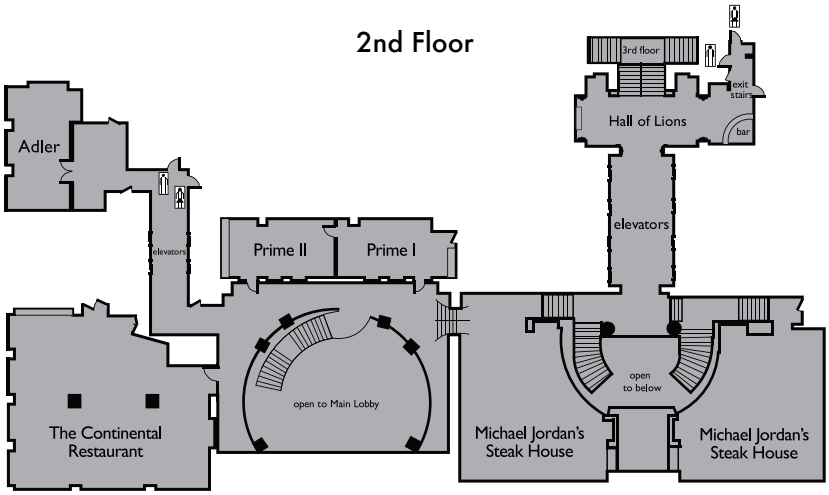
Beyza Aksu
Lokman Akbay
Diego Luna Bazaldua
Lisa Beymer
Jeremy Brown
Tianna Chantel Floyd
Alejandra Garcia
Jerusha Gerstner
Jason Herron
Xueying Hu
Hong Jiao
Taeyoung Kim
David King
Naama Lewis
Jin Liu
Matthew Madison
Ashley Sandoval
Can Shao
Ben Shear
Phil Sherlock, Jr.
Ah Young Shin
Lauren Stevenson
Joshua Sussman
Ragip Terzi
Bing Tong
Danielle Tyree
Tina Wang
Dawn Woods
Ping Yang
Jing-Ru Xu
Nedim Yel

Future Annual Meeting

2016 Annual Meeting
April 7 - 11
Washington, DC

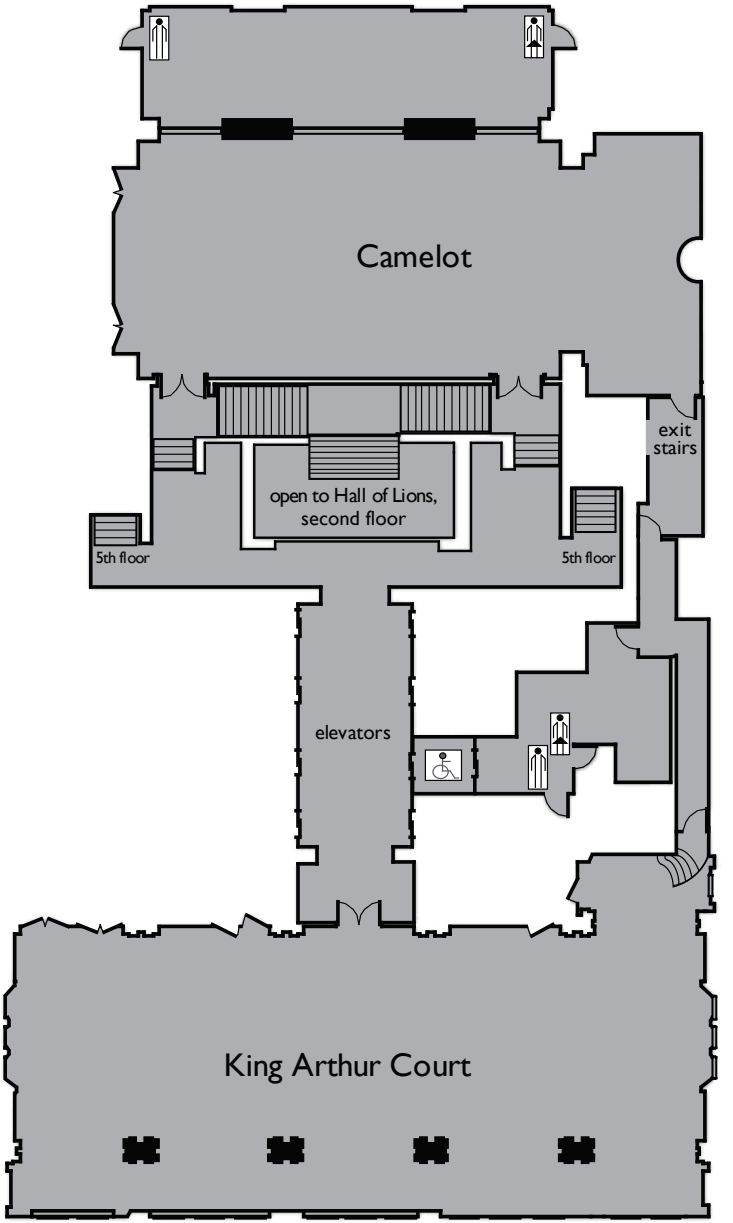
Hotel Floor Plans – InterContinental Hotel Chicago





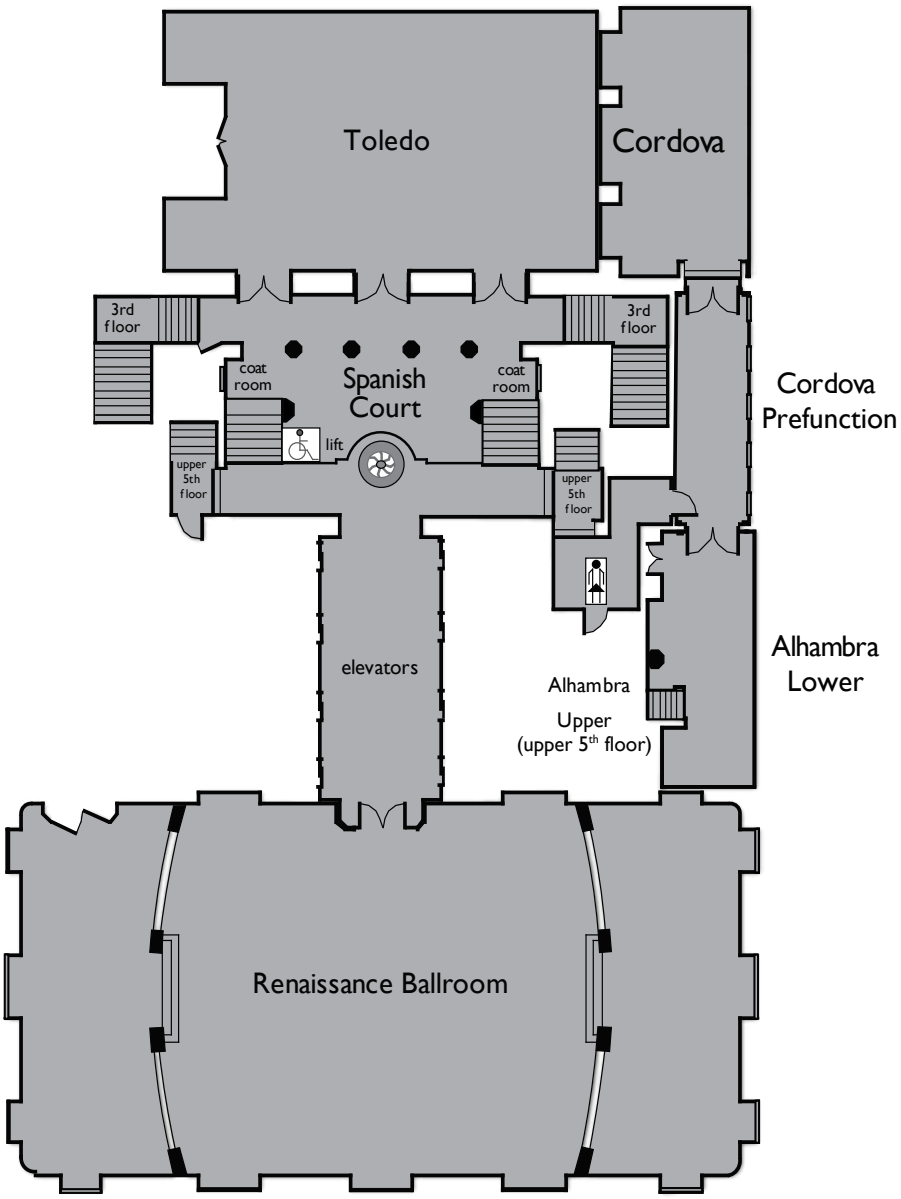
Hotel Floor Plans – InterContinental Hotel Chicago

3rd Floor



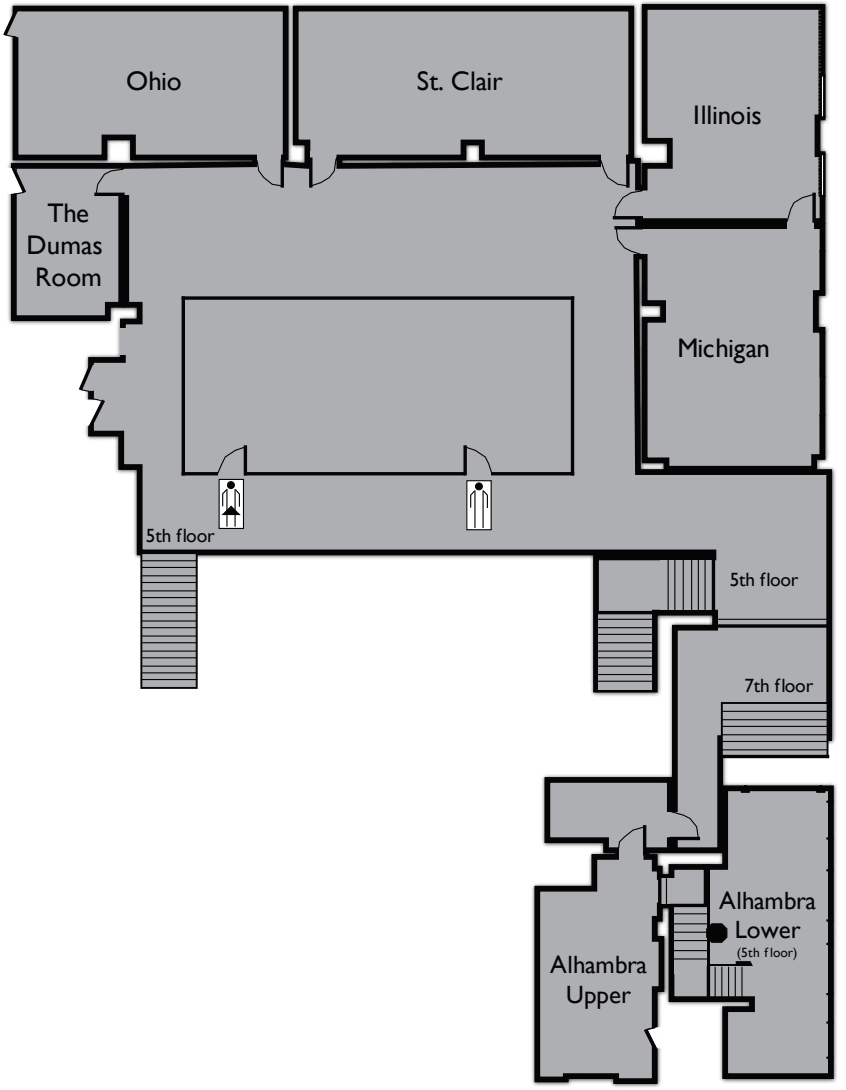
Hotel Floor Plans – InterContinental Hotel Chicago

5th Floor



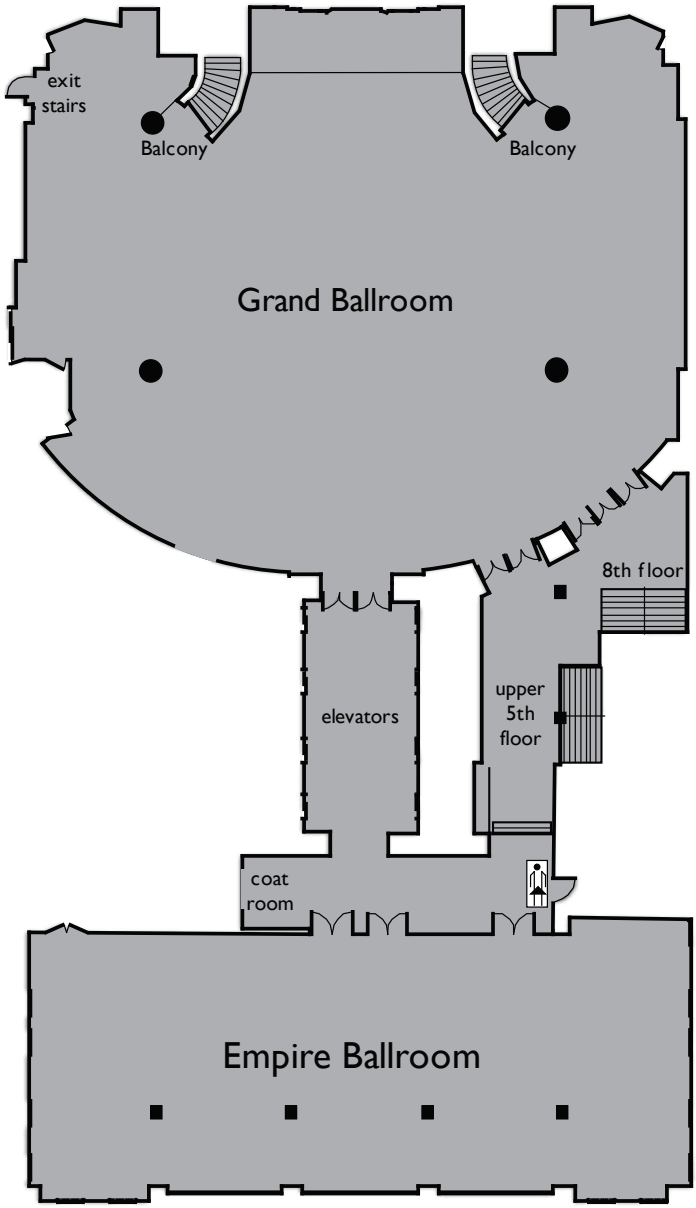
Hotel Floor Plans – InterContinental Hotel Chicago

Upper 5th Floor



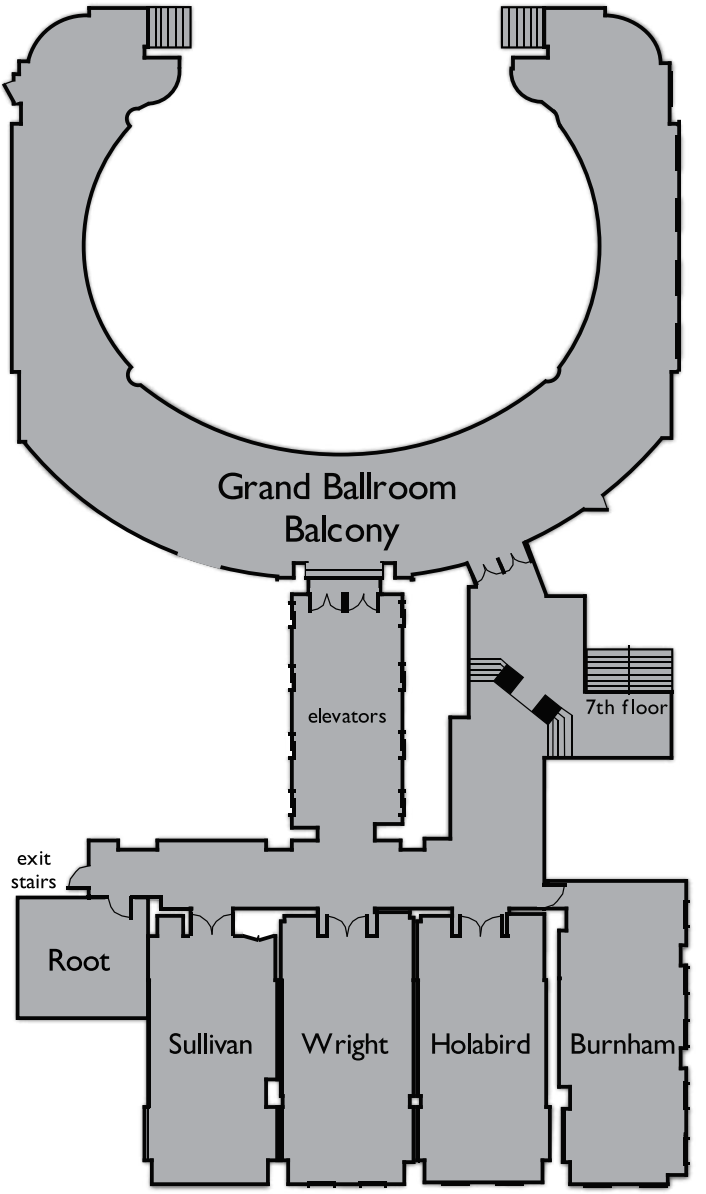
Hotel Floor Plans – InterContinental Hotel Chicago

7th Floor



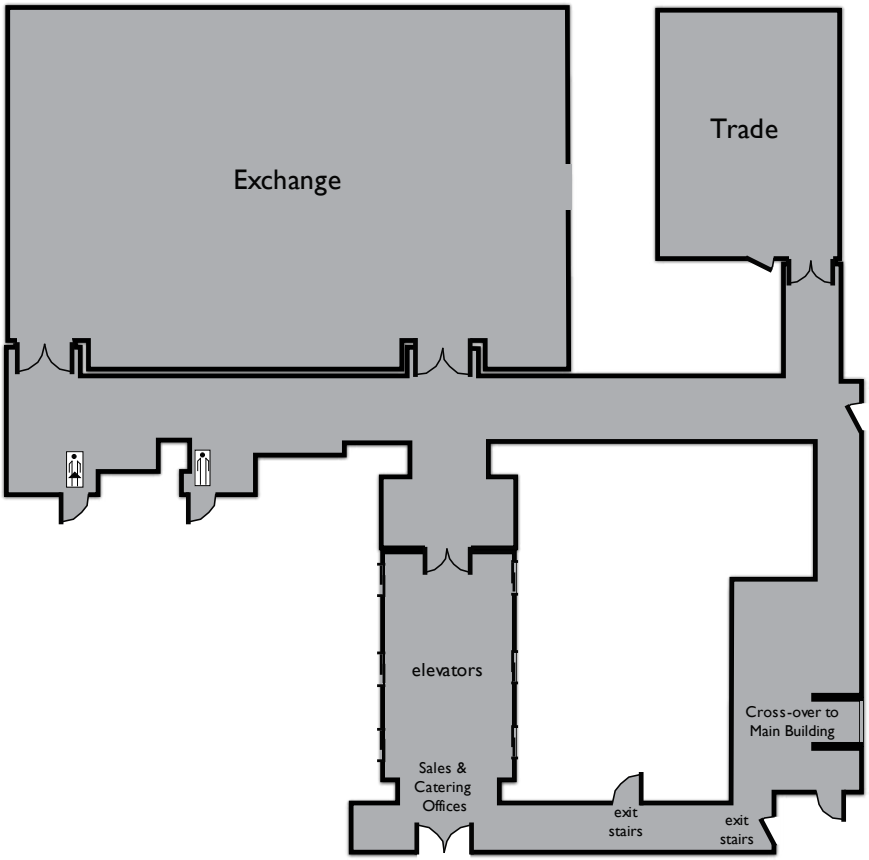
Hotel Floor Plans – InterContinental Hotel Chicago

8th Floor



Hotel Floor Plans – InterContinental Hotel Chicago

11th Floor



Pre-Conference Training Sessions

The 2015 NCME Pre-Conference Training Sessions will be held at the InterContinental Hotel on Wednesday, April 15, and Thursday, April 16. 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 noon. All half-day afternoon sessions will run from 1:00 PM to 5:00 PM.

On-site registration for the Pre-Conference Training Sessions will be available at the NCME Information Desk at the InterContinental Hotel 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 training sessions that have pre-paid an additional fee.

Pre-Conference Training Sessions - Wednesday, April 15, 2015

Wednesday April 15, 2015
8:00 AM - 12:00 PM, Exchange, 11th Floor, Training Session, AA

A Practitioner's Guide to Growth Models

Katherine Furgol Castellano, Educational Testing Service and Andrew Ho, Harvard Graduate School of Education

Practitioners use growth models to support inferences about student learning, educator effectiveness, and large-scale educational progress. In educational accountability systems, growth models have become increasingly complex, combining statistical models with calculations motivated by policy decisions. As the stakes on growth models rise, so does the importance of understanding their intricacies.

This training session reviews and compares seven popular growth models, including gain-based models, categorical models, projection models, and Student Growth Percentiles, by answering six critical questions for each model. These questions help to identify, for example, the primary interpretations each growth model supports, the data requirements of each model, and possible unintended consequences of using each model in an accountability system.

By the end of the session, participants should be able to articulate contrasts between popular growth models as well as actively compare growth model results using real datasets in Excel and/or R.

Wednesday April 15, 2015
8:00 AM - 12:00 PM, Seville Ballroom West, Lobby Level, Training Session, BB

An Introduction to Equating in R

Anthony Albano, University of Nebraska-Lincoln, Lincoln, NE

This training session introduces participants to observed-score and item response theory (IRT) equating methods through a series of exercises involving analysis of real data within the statistical environment R. Researchers and practitioners are invited to participate. A background in introductory statistics and experience using R are recommended but not required.

Many testing programs collect data on multiple forms administered across time and/or across different samples of individuals. These programs include large-scale applications, such as in licensure and admissions testing, and smaller-scale applications, such as in classroom assessment and intervention studies. In each case, practitioners and researchers can utilize equating procedures to convert multiple test forms to a common measurement scale.

Experience has shown that individuals tasked with equating often lack the education and training required to do so. The misuse of equating procedures can result in invalid score interpretations. This session provides participants with a brief and practical induction to equating principles and concepts and to the procedures needed to effectively use equating. The session begins with an introduction to R and to observed-score equating and IRT methods. The majority of the session is then devoted to a series of exercises requiring participants to prepare and analyze provided data from a variety of test administration designs. These exercises address presmoothing and equating using observed-score methods, equating/linking using IRT methods, and visualizing, summarizing, and evaluating results.

A background in introductory statistics and experience using R are recommended but not required. Participants should bring their own computers, with R (R Core Team, 2014) and the most recent version of the equate package (Albano, 2014) installed. Electronic training materials will be provided via email at least one week prior to the conference.

Wednesday April 15, 2015
8:00 AM - 12:00 PM, St. Clair, Upper 5th Floor, Training Session, CC

Using Visual Displays to Inform Assessment Development and Validation

Brett Foley, Alpine Testing Solutions, Denton, NE

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). Presenters will provide step-by-step instruction on how to create the various displays using readily available software. Participants should bring a laptop or similar device loaded with Microsoft Excel (2010 version highly recommended). Panelists will receive flash drives with Excel files and instructions for creating and adapting the visuals discussed in the workshop.

With any session involving technology integration, there is a tendency to overload participants with software features. To respond to this challenge, presenters will provide some illustrations, but intersperse the hands-on opportunities to discussion of visual displays principles to allow for greater depth of participation by participants; panelists will also be given videos providing instruction for each activity for later reference and review.

Objectives are to provide assessment developers, users, and consumers (a) relevant examples of visual data displays designed to facilitate test development and validation processes (e.g., program design, content specification, item writing, item review, standard setting, score reporting) and (b) experience creating such displays.

Wednesday April 15, 2015
8:00 AM - 5:00 PM, Empire Ballroom, 7th Floor, Training Session, DD

Leveraging Open Source Software and Tools for Statistics/Measurement Research

Damian Betebenner, Center for Assessment; Adam VanIwaarden, and Ruhan Circi, University of Colorado, Boulder

Measurement and statistics specialists have used software for decades with tools like SPSS, SAS, and Stata and more recently the open source software environment R. The expansion of the importance of software goes well beyond software packages data analysts use. Development tools alter the way that people work, collaborate and disseminate the results of their efforts. This training session will introduce users to the rapidly expanding universe of open source tools available that can be used to increase the transparency and reproducibility of their research while simultaneously enhancing productivity, collaboration and dissemination.

In this full-day session, participants will be introduced to open modern software analysis and development tools and show how, through rich, real-life working examples, they can be combined to enhance the goal of producing transparent and reproducible research. Example projects will be presented to participants that range from a dissertation, to a prototype for a published article, to a multi-state/national data analysis project.

Wednesday April 15, 2015
8:00 AM - 5:00 PM, Renaissance Ballroom, 5th Floor, Training Session, EE

flexMIRT®: Flexible Multilevel Multidimensional Item Analysis and Test Scoring

Li Cai, UCLA and Carrie Houts, Vector Psychometric Group, LLC

There has been a tremendous amount of progress in item response theory (IRT) in the past two decades, resulting in interesting new software implementations for research and operational use. flexMIRT; is an IRT software package which offers multilevel, multidimensional, and multiple group item response models. flexMIRT also offers users the ability to obtain recently developed model fit indices, fit diagnostic classification models and models with non-normal latent densities. This training session is intended to provide a broad overview of the features of flexMIRT; as well as hands on experience using the software. Attendees will receive a free two-month trial version of flexMIRT. It is assumed that attendees will be familiar with IRT. It would be helpful if the attendees could bring their own devices running Windows 7 or above.

flexMIRT fits a variety of unidimensional and multidimensional IRT models as well as extended diagnostic classification models, to single-level and multilevel data using maximum marginal likelihood (or optionally modal Bayes) estimation. It produces IRT scale scores using maximum likelihood (ML), maximum a posteriori (MAP), and expected a posteriori (EAP) estimation. It (optionally) produces summed-score to IRT scale score (EAP) conversion tables for single-level IRT models. As for the item types, flexMIRT; can estimate any combination of 3-parameter logistic (3PL) model, logistic graded response model (which includes 2PL and 1PL as special cases), and the nominal categories model (including any of its restricted sub-models such as generalized partial credit model, partial credit model, and rating scale model) for both single-level and multilevel data, in any number of groups. The availability of generalized dimension reduction EM algorithm as well as the Metropolis-Hastings Robbins-Monro (MH-RM) algorithms, coupled with arbitrary user-defined parameter constraints, make flexMIRT; one of the most flexible IRT software programs either commercially or freely available today.

flexMIRT also has some of the richest psychometric and statistical features. flexMIRT supports several methods for estimating item parameter standard errors. A multitude of model fit statistics for dimensionality analysis, item-fit testing, and latent variable normality diagnosis are included in flexMIRT. Its multiple-group estimation features easily facilitate studies involving differential item function (DIF) and test linking (including vertical scaling).

Another innovation in flexMIRT is its ability to relax the ubiquitous multivariate normality assumption made in virtually all IRT models. With an extended dimension reduction algorithm, it supports the non-parametric estimation of latent density shapes using empirical histograms for both unidimensional and hierarchical (e.g., bifactor and testlet response theory) item factor models, and in any number of groups, with support for constraints on group means and variances. This feature is also fully integrated into the built-in Monte Carlo simulation module that can generate data from any model implemented in flexMIRT.

Windows-based flexMIRT has an intuitive syntax and friendly graphical user interface (GUI), available in both 32-bit and 64-bit flavors. A newly-designed memory allocation scheme helps flexMIRT efficiently handle thousands of items and millions of respondents, with no imposed upper limit on the size of the problem.

Wednesday April 15, 2015
8:00 AM - 5:00 PM, Seville Ballroom East, Lobby Level, Paper Session, FF

An Introduction to Diagnostic Classification Modeling

Laine Bradshaw, University of Georgia, Athens, GA

Diagnostic classification models (DCMs) can efficiently provide reliable feedback from multidimensional tests. First, this workshop provides a semi-technical introduction to the terms and techniques used for diagnosing what students know. Then, participants will gain hands-on experience estimating and interpreting DCMs using software provided for participants own laptops.

Upon completion of the workshop, participants will be able to understand the rationale and motivation for using diagnostic classification models. Furthermore, participants will be able to understand the types of data typically used in diagnostic measurement along with the information that can be obtained from implementing diagnostic models. Participants will become well-versed in the state-of-the-art techniques currently used in practice and will be able to use and estimate diagnostic measurement models on their own.

From a practical point-of-view, participants will see how to develop instruments for diagnosing student abilities and how to create easy-to-use score reports. Additionally, participants will be able to interpret results from diagnostic measurement analyses to evaluate student mastery profiles and understand how to use profiles to inform instructional plans that focus on a multidimensional view of student progress in achievement. Finally, participants will be able to interpret research articles using diagnostic measurement techniques, thereby allowing students a better opportunity to integrate such methods into their active research programs.

Wednesday April 15, 2015
8:00 AM - 5:00 PM, Toledo, 5th Floor, Training Session, GG

Optimal Test Design

Wim van der Linden, Qi Diao and Jie Li, CTB McGraw-Hill Education, Monterey, CA

The topic of IRT-based test assembly was introduced by Birnbaum's in his contribution to the well-known book by Lord and Novick (1969). Its basic idea exists of the assembly of a test form to have an information function matching a target function for the intended application.

Although intuitively convincing, Birnbaum's procedure was not practical yet. In practice, test forms are never assembled to meet a statistical target only; they always have to meet a potentially large variety of other specifications as well, for instance, blueprints for its content, certain answer key distributions, a given time slot, exclusion of specific combinations of items, or bounds on their total word count. In fact, as will be demonstrated by our examples in the training session, it is not unusual for real-world test-assembly problems to involve hundreds of additional constraints on the selection of the items. It is not immediately clear how to meet each of them while manipulating a test-information function with respect to a target as suggested by Birnbaum.

But even without any of these constraints, the job of picking an optimal combination of items is already impossible. The reason is an instance of the well-known combinatorial explosion. The number of possible different test forms of length n from a pool of I items is equal to (I n) generally a prohibitively large number. For instance, even for a pool of only I=50 items, the number of different forms of n=10 items is already much greater than the current world population. Only methods with mathematically proven optimality are able to deal with such explosions of possibilities; we will never know if a solution proposed by a method based on a heuristic idea, or just by manual selection, will be the best available from the item bank.

Practical test-assembly methods are even more complicated in that we hardly ever assemble just one single form consisting of discrete items at a time. Often, we need a set of forms required to be completely parallel, possibly with constraints on the item overlap between some or all of them. Or a set that addresses systematical differences in expected achievements between groups of test takers. Or the item pool may consist of items organized around common stimuli and we have to impose bounds on the numbers of items selected per stimulus. Even when assembling one form at a time, it seems prudent to keep an eye on the assembly of all future forms. In order to avoid deterioration of test quality over time we may already want to compromise between what is taken from and left in the pool early on.

Finally, real-world testing programs with frequently assembled forms tailored to specific applications that are delivered electronically generally require fast algorithms to produce them. Ideally, we should be able to assemble them in real time. The requirement of real-time solutions even becomes mandatory when tests are assembled adaptively rather than as fixed forms.

The goal of this training session is to show that all these problems can be solved by treating test assembly as an instance of combinatorial optimization. The basic methodology exists of translating all test specifications in a set of constraints with an objective function, model the objective function and constraints using binary decision variables, and having a standard mathematical solver find the solution to the optimization problem (van der Linden, 2005).

...continued on page 22

...continued from page 21

The first lecture part of the workshop introduces the principles of item-response theory (IRT) required for test development, reflects on the history of test design, explains Birnbaum's approach to IRT-based test assembly, and shows how his approach can be implemented to be useful for nearly every practical form of test assembly using the methodology of constrained combinatorial optimization. The second lecture introduces the core methodology and shows how every content, statistical, and practical test specification can be modeled as an objective for or constraint on item selection from an IRT-calibrated item pool. The third lecture demonstrates the use of the methodology for the selection of a variety of single-form test assembly problems as well as problems of simultaneous selection of multiple forms that have to be parallel is systematically different in content and/or statistical characteristics. All models will be illustrated with examples from real-world testing programs. The final lecture addresses the topic of adaptive testing as a special version of optimal constrained test assembly implemented through a shadow-test approach. It shows how the approach can be used to implement every form of linear-on-the-fly, multistage, or adaptive testing and discusses how such problems as content balancing, item exposure control, and control of differential speededness can be solved just by selecting the right combination of constraints for the test-assembly model.

In addition to the lectures, participants will be offered demos of the Optimal Test Assembler™ and ShadowCAT™ programs and work with lpSolveAPI for R on the test-assembly problems presented in Diao and van der Linden (2011) during computer exercises.

Each of the instructors is specialized in constrained combinatorial optimization and test assembly and has ample experience with applications of its methodology to practical testing problems. They have also provided leadership to the team that developed the software demonstrated during the session.

Wednesday April 15, 2015
1:00 PM - 5:00 PM, Exchange, 11th Floor, Training Session, HH

An Overview of Operational Psychometric Work in Real World

JongPil Kim, ACT; Laura Kramer, University of Kansas; Jinghua Liu, SSATB; Hyeonjoo Oh, Educational Testing Service; Leslie Keng, and Ye Tong, Pearson

The purpose of this training session is to provide an overview of psychometric work that is routinely performed by testing organizations. The work scope includes evaluation of items and test forms written and assembled by test development specialists, item analysis and test analysis, equating and scaling, score reporting, field test design, standard setting, etc. These statistical activities are conducted with specific purposes of ensuring the quality of a testing program, reported scores and supporting appropriate interpretations of these scores. This training session describes the interpretation and communication of analysis results to test score users as well. This training session will focus on four topics: (1) outline of operational psychometric activities across different testing companies, (2) hands-on activities related to item review and test form review, (3) hands-on activities related to reviewing and interpreting equating output and making decisions, and (4) discussion session regarding factors that affect operational psychometric activities such as testing mode comparability. The current training facilitates various professional psychometric skills and research knowledge, as well as describes the applications of recent methodological developments adopted in practice. Hands-on examples and activities will also be included as part of the training session to provide the participants some real world examples. Representatives from different testing organizations and University research center will present various topics related to processes in an operational cycle.

Many graduate students and junior level psychometricians have knowledge on psychometrics and measurement theory but may not have enough opportunities to expose themselves to the real world psychometrics work. We are hoping that through this training session, participants will get a glimpse of the entire operational cycle, as well as gain some understanding of the challenges and practical constraints that psychometricians face at testing organizations. After the training, we are expecting participants are able to evaluate item analysis and equating results. Each presenter will directly interact with training participants and work together with them. Questions and answers will be encouraged and entertained at any point during the session.

Wednesday April 15, 2015
1:00 PM - 5:00 PM, Seville Ballroom West, Lobby Level, Training Session, II

A Graphical and Nonlinear Mixed Model Approach to IRT with the R Package Flirt

Frank Rijmen and Minjeong Jeon, OSU

The first goal of the workshop is to show how generalized linear and nonlinear mixed models offer a powerful statistical framework for item response theory models. Ability dimensions in item response theory models are conceptualized as random effects in the mixed model framework, and the responses to items correspond to repeated measurements of the same individual. Random effects are unobserved or latent variables that correspond to sources of individual differences. They account for the dependencies that are typically observed among responses clustered within the same person. The advantages of working within this overarching framework are substantial. First, the common framework helps to understand the commonalities and differences between various item response theory models. Second, models can be extended, at least conceptually, in a straightforward way. Third, theoretical and empirical findings can be more easily communicated with a larger research community through the use of a common terminology.

The second goal of the workshop is to show how the parameters of multidimensional item response theory models can be estimated with an efficient EM algorithm that is embedded within a graphical model framework. Maximum likelihood estimation of model parameters in generalized linear and nonlinear mixed models involves integration over the space of all random effects. In general, the integrals have no closed-form solution. Numerical integration over the joint space of all latent variables becomes computationally very demanding as the number of dimensions grows. This technical challenge has hampered the use of multidimensional item response theory in operational settings. However, depending on the conditional independence relations between the dimensions one is willing to assume, the actual computational cost can be lowered by exploiting these conditional relations during parameter estimation. In particular, the set of conditional independence relations implied by a model can be used to partition the joint space of all latent variables into smaller subsets that are conditionally independent. As a consequence, numerical integration by enumeration over the joint latent space can be replaced by a sequence of integrations over smaller subsets of latent variables. The gain in efficiency can be dramatic in some cases. Graphical model theory offers a general procedure for exploiting conditional independence relations during parameter estimation.

Thirdly, we will present the recently developed R package flirt (flexible item response theory modeling). The package relies on an integration of generalized linear and nonlinear mixed models on the one hand, and graphical models on the other hand. As a result, it is more general and efficient than other existing R packages for item response theory models. The participants will have the opportunity to familiarize themselves with the flirt package during various hands-on sessions throughout the workshop.

Wednesday April 15, 2015
1:00 PM - 5:00 PM, St. Clair, Upper 5th Floor, Training Session, JJ

Cognitive Lab Techniques: An Overview, Framework, and Some Practice

Irvin R. Katz, Jung Aa Moon, and Teresa King, Educational Testing Service, Princeton, NJ

Cognitive labs have become increasingly popular over the past decades as methods for gathering detailed data on the processes by which test-takers understand and solve assessment items and tasks. Cognitive labs result in data that may inform a wide variety of research and practical issues in the field of educational measurement, ranging from assessment development (e.g., "are test takers confused by the phrasing of this item?") to validity (e.g., "does this item engage the expected knowledge and skills from test takers?"). For example, within validity, cognitive labs can provide evidence on response process validity (AERA, APA, & NCME, 1999), help detect construct-irrelevant factors affecting test-taker performance, and address various fairness and accessibility issues.

Nonetheless, even a quick review of the Procedures section of a few cognitive lab studies reveals that the phrase "cognitive labs" is used to describe different techniques (Arieli-Attali et al., 2011): think aloud or verbal reports as in Ericsson and Simon's (1984) methodology (e.g., Barkaoui, 2011; Baxter, & Glaser, 1998; Katz, 1994), cognitive interview and/or retrospective verbalization (e.g., Almond et al. 2009; Hansen, 2009; Snow & Katz, 2009), stimulated retrospective (Feng & Sand, 2013), and others. How are researchers or practitioners to know what techniques yield the right type of data to address particular research or practical questions? A poorly designed cognitive lab can yield data that do not help address the question of interest or, worse yet, lead to unwarranted conclusions (cf. Leighton, 2004).

By the end of the workshop, attendees should have a greater appreciation for the range and variety of cognitive lab techniques as well as a framework for organizing the techniques in terms of their related research questions. This understanding should aid attendees in their own work on conducting cognitive labs, in interpreting data from cognitive labs, and in critically reading research literature that utilize cognitive lab techniques.

Pre-Conference Training Sessions - Thursday, April 16, 2015

Thursday April 16, 2015
8:00 AM - 12:00 PM, Exchange, 11th Floor, Training Session, KK

Fundamentals of Item Response Theory and Computerized Adaptive Testing

David J. Weiss, University of Minnesota and Alper Sahin, Cankaya University

Computerized adaptive tests (CAT) have had substantial impact over the past decade, especially after they were put into use by state-led consortia (e.g. Smarter Balanced, WIDA) for accountability purposes. The rise of CAT in educational measurement has manifested itself in a need to train educational researchers, practitioners, and educational managers on Item Response Theory (IRT) and CAT.

This training session will address the basics and fundamentals of IRT and CAT. It will provide the participants with a broad overview of what IRT and CAT are and how they can be implemented by educational institutions. In addition, attendees will have some practical demonstrations on how to use some specialized IRT and CAT software (e.g. CATsim and Xcalibre). Participants will be actively involved in the training through group discussions and some practice activities. They will also be provided with electronic copies of the materials including the Powerpoint presentations and related handouts used in the session. The intended audience includes graduate students, multi-field educational researchers, practitioners, and educational management professionals who have little or no knowledge of IRT and CAT.

Upon completion of this training session, participants will be able to build the prerequisite knowledge base for developing their own IRT-based research or CATs and develop an understanding of how the assessment systems of the 21st Century function. In parallel with this, the participants will also develop the rationale and motivation to use CAT for accountability purposes.

Thursday April 16, 2015
8:00 AM - 12:00 PM, Seville Ballroom East, Lobby Level, Training Session, LL

Item Response Theory With jMetrik and Psychometric Programming With Java

Patrick Meyer, University of Virginia

jMetrik is an open source program for psychometrics. It is a user-friendly program that incorporates a common data source and a variety of procedures for measurement such as tools for classical test theory, item response theory, scale linking and score equating. Among the psychometric procedures are marginal maximum likelihood estimation of item parameters in the three parameter logistic model (3PLM) and generalized partial credit mode (GPCM). As a pure Java application, jMetrik runs on windows, Linux, or Mac OSX operating systems using either 32- or 64-bit processors. This workshop teaches participant to use jMetrik for applications of item response theory. It also introduces them to programming with Java and the source code that drives jMetrik.

In the first part of the workshop, participants will use jMetrik to analyze test data. They will use jMetrik to estimate item and person parameters for the 3PLM and GPCM. Participants will also learn to create various plots such as item characteristic curves and information functions.

jMetrik is built entirely with the Java programming language. Source code is divided into two main libraries: jMetrik and psychometrics. The jMetrik library provides the interface and database related functionality, while the psychometrics library contains the measurement-related code such as classes for item response models and MMLE. Working with these libraries requires an understanding of object oriented programming and the Java language.

Many people are familiar with statistical and psychometric programming with R, SAS, and other software, but they are less familiar with lower-level languages such as Java. In the second part of this workshop, participants will be introduced to object oriented programming with Java and how to code, compile and execute their own program. They will then learn about the psychometrics library and how to use it for item response theory, building their own application or adding functionality to the library.

Thursday April 16, 2015
8:00 AM - 12:00 PM, Seville Ballroom West, Lobby Level, Training Session, MM

Landing Your Dream Job for Graduate Students

Deborah Harris, Nathan Wall, and Xin Li, ACT

This training session will address practical topics graduate students in measurement are interested in regarding finding a job and starting a career, concentrating 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, including networking, internships, and volunteering, 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), and the interview process (job talks, questions to ask, 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).

Thursday April 16, 2015
8:00 AM - 5:00 PM, Empire Ballroom, 7th Floor, Training Session, OO

Multidimensional Item Response Theory: Theory and Applications and Software

Lihua Yao, DMDC, Mark Reckase, Michigan State University, and Rich Schwarz, Educational Testing Service

Theories and applications of multidimensional item response theory model (MIRT) and Multidimensional Computer Adaptive testing (MCAT) and MIRT linking are discussed. Software BMIRT, LinkMIRT, SimuMIRT, and SIMUMCAT are demonstrated. BMIRT (Yao, 2003) is a computer program that estimates item and ability parameters in the multidimensional multi-group IRT framework; exploratory and confirmatory approaches are supported. LinkMIRT (Yao, 2004) is linking software that links two sets of item parameters onto the same scale in the MIRT frame work. SimuMIRT is software that simulates data for various MIRT models. SimuMCAT (Yao, 2011) is a computer program for MCAT simulation, which has five MCAT item selection procedures with item exposure control methods and content constraints. This session is intended for researchers who are interested in learning and understanding MIRT, MIRT linking, and MCAT and their applications and who are working with dichotomous or polytomous data that is multidimensional in nature. BMIRT supports the three-parameter logistic model, generalized two-parameter partial credit model, graded-response, rater model, and testlet-effect models.

Participants should bring laptop computers and any data they would like to use. Participants are required to download the free software to the laptop they intend to use prior to the session at www.BMIRT.com for the hands on experience. Participants are required to download the Java Runtime Environment or JRE to the laptop computer.

This workshop will introduce MIRT, MIRT linking theory, and MCAT and conduct hands-on experience using BMIRT, LinkMIRT , SimuMIRT, and SIMUMCAT. There are demonstrations and discussions of results and output. Extensive uses of representative examples included with the program are used to guide learning. Hands-on activities are 50% of the session.

The participants will (a) learn the associated concepts and gain a comprehensive understanding of MIRT, MIRT linking, and Multidimensional CAT (b) know the applications of MIRT, MIRT linking, and MCAT, (c) understand appropriate uses of BMIRT, LinkMIRT, SimuMIRT, and SIMUMCAT (d) understand the data input requirements and formats, and (e) understand and be able to interpret the output files.

Thursday April 16, 2015
8:00 AM - 5:00 PM, Renaissance Ballroom, 5th Floor, Training Session, PP

Generalizability Theory and Applications

Robert Brennan and Won-Chan Lee, University of Iowa, Iowa City, IA

Generalizability theory liberalizes and extends classical test theory. In particular, generalizability theory enables an investigator to disentangle multiple sources of error through the application of analysis of variance procedures to assess the dependability of measurements.

The primary goals of this training session are to enable participants to understand the basic principles of generalizability theory, to conduct relatively straightforward generalizability analyses, and to interpret and use the results of such analyses. Mathematical and statistical foundations will be treated only minimally. Major emphasis will be placed upon quickly enabling participants to conduct and interpret relatively straightforward generalizability analyses, then more complicated ones. Examples will include various types of performance assessments.

In general, the goals of this training session are to enable participants to understand the basic principles of generalizability theory, to conduct G(eneralizability) studies and D(ecision) studies, and to interpret and use the results of generalizability analyses. More specifically, the objectives of this training session are as follows:

(A) Basic Principles
(1) An understanding of principal similarities and differences between generalizability theory and other psychometric theories;
(2) An understanding of principal similarities and differences between generalizability theory and analysis of variance;
(3) Knowledge and understanding of the fundamental concepts in generalizability theory, including universe of admissible observations, universe of generalization, G studies and D studies, facets and objects of measurement, variance components, universe score, variance, error variances, and generalizability coefficients.

(B) Computational and Estimation Procedures
(4) Ability to compute by hand, or with a hand calculator, estimates of variance components and other parameters in generalizability theory for simple designs, given mean squares;
(5) Ability to employ computer programs such as GENOVA to estimate parameters for real data sets and balanced designs;
(6) Knowledge of the characteristics of estimates of variance components, and other parameters in generalizability theory.

(C) Applications
(7) Ability to appropriately interpret results from generalizability analyses;
(8) Ability to design reasonable and useful G and D studies;
(9) Ability to conduct generalizability analyses with not-too-complicated real data sets;
(10) An understanding of appropriate and inappropriate uses of generalizability theory.

All attendees are asked to bring a laptop computer.

Thursday April 16, 2015
8:00 AM - 5:00 PM, Toledo, 5th Floor, Training Session, QQ

Bayesian Networks in Educational Assessment

Duanli Yan, Educational Testing Service; Russell Almond, Florida State University; Robert Mislevy, and David Williamson, Educational Testing Service

The Bayesian paradigm provides a convenient mathematical system for reasoning about evidence. Bayesian networks provide a graphical language for describing and reasoning in complex models. This allows assessment designers to build scoring that have fidelity to cognitive theories, are mathematically tractable and can be refined with data. (Book is included).

This course will provide the background information on Bayesian networks, Graphical Models and related inference and representation methods and provide examples of their use in educational assessment. Although the course will review the Evidence Centered Design framework for representing measurement models in educational assessments using graphs, the primary goal is to review the work done in other communities for psychometricians and psychologists.

Then, after a brief overview of the most commonly used Bayesian network tools, it will provide a well-received interactive hands-on session on using Bayesian network tool on small examples for Bayesian inference, manipulating graphical models and applications in educational assessment. It will also review the existing body of literature on graphical models from other disciplines (in particular, the Uncertainty in Artificial Intelligence literature).

The course will consist of a series of lectures interspersed with formative assessment, and hands-on examples. The live examples will provide illustrative examples of graphical models in education with some live computations.

Topics covered are evidence-centered assessment design, basic Bayesian network representations and computations, available software for manipulating Bayesian networks, refining Bayesian networks using data, and example systems using Bayesian networks. The last application was the focus of the presenter's 2000 NCME Award for Outstanding Scientific or Technical Contribution to Educational Measurement.

The training course consists of the following 5 sessions:

1. Evidence Centered Design
2. Graphical Models
3. Graphical Modeling Tools and Applications
4. Refining Graphical Models With Data
5. ACED: ECD in Action Demonstration.

Thursday April 16, 2015
1:00 PM - 5:00 PM, Exchange, 11th Floor, Training Session, RR

Advances in Measuring 21st Century Skills: Constructs, Development, and Scoring

Patrick Kyllonen and Jonas Bertling, Educational Testing Service, Princeton, NJ

The workshop begins with an overview of the issues involved in assessing noncognitive skills. We identify a number of frameworks, such as 21st century skills (e.g., ATC21S, NRC), and school frameworks (e.g., University of Chicago Consortium on Chicago School Research, the California Office to Reform Education, the INDEX Mission Skills Assessment), and discuss overlaps with the five-factor personality model. We review methods for assessing noncognitive skills: self-, teacher, and peer assessments, anchoring vignettes, single and multiple dimensional forced choice and ranking methods, situational judgment tests, and implicit association and conditional reasoning tests. We review classical and IRT-based scoring methods for each assessment, and discuss pros and cons. We also discuss score reporting, to highlight current status, comparisons, and trends, and review approaches for collecting and analyzing data on background variables.

We examine the use of certain methods in detail, including anchoring vignettes, forced-choice and ranking methods, and situational judgment tests. For anchoring vignettes we review writing from construct and item definitions, administering, scoring (primarily nonparametric scoring), and reporting. We review what is known and discuss item development procedures starting from construct definitions and using those to create vignettes at various levels on the trait continuum (e.g., low, medium, and high skill). We also discuss practical matters such as whether to administer them before or after items, and with or separately from items. We discuss scoring anchoring vignettes, and provide hands-on practice. Participants will be shown how to write code to score anchoring vignette tasks. We review item analysis techniques and discuss reporting on vignette-adjusted scores.

For forced-choice methods we review approaches (e.g., pairs, sets of 3, sets of 4; asking for more like; least like; and both, other ranking methods), and review advantages and disadvantages of ipsative and quasi-ipsative scoring approaches, using both classical and IRT. We review findings to help participants understand when one might use forced-choice approaches.

For situational judgment testing, we review item development, including approaches for collecting critical incident data from experts, and transforming such data into test items. We compare testlet- and single statement rating scale items, and multiple-choice items, both and; select-the-best; and select-the-worst; and select what you would do formats. Participants receive hands-on experience in developing SJT items. Both classical and item-response theory scoring approaches are reviewed, in particular the nominal response model for scoring, along with visualization plots for determining appropriate scoring models. We review issues in reporting.

Thursday April 16, 2015
1:00 PM - 5:00 PM, Seville Ballroom East, Lobby Level, Training Session, SS

Using IRT for Standard Setting in Performance Based Assessments

Boaz Shulruf and Philip Jones, University of New South Wales, Sydney, Australia

Item Response Theory and related models are complex statistical techniques, mostly used for improving test and item quality with limited application in standard setting. This workshop demonstrates how the Rasch model could be used for standard setting in performance based assessments including the advantages and limitations of this approach.

IRT models utilize the information in tests to estimate person and items parameters. This workshop will focus on applying IRT, particularly the Rasch model, for setting standards in educational assessment. Most of the currently used standard setting methods rely upon panels of judges to make decisions on the expected performance of a minimally competent examinee in a given test, either at the whole test or at an item level. Some methods use information obtained from IRT analyses to support judges' decisions (Clauser, Mee, Baldwin, Margolis, & Dillon, 2009; Wang, 2003).

The practical challenge in standard setting is to make a defensible decision on scores that are neither clearly pass nor clearly fail, hence borderline. Ben-David phrased it well by defining the purpose of standard setting to "separate the non-competent from the competent” (Ben-David, 2000, p. 120). This challenge is particularly difficult in examinations that use grading systems that include the category or score range of a 'borderline performance' (e.g. clinical examinations, open-ended type of examination). However, if the main purpose of the assessment is to distinguish between competence and incompetence, a borderline grade fails to do so (Boursicot, Roberts, & Pell, 2007; Roberts, Newble, Jolly, Reed, & Hampton, 2006; Schoonheim-Klein et al., 2009; Wilkinson, Newble, & Frampton, 2001).

To address this critical issue, a myriad of standard setting methods have been developed, aiming to identify a cut-score on a continuous scale that best distinguishes between competent and incompetent examinees (Cizek & Bunch, 2007; Kane, 2013; Zieky, 2012). Nonetheless, despite this broad range of methods, concerns about reliability, validity and acceptability of these cut-scores remain due particularly to unavoidable bias related to judges' decisions (Chang, Dziuban, Hynes, & Olson, 1996; Wayne, Cohen, Makoul, & McGaghie, 2008).

Improving the reliability of judge-based methods normally requires recruiting a large number of experts who need to go through a lengthy and expensive process (Ben-David, 2000; Brannick, Erol-Korkmaz, & Prewett, 2011; Cizek & Bunch, 2007; Hurtz & Auerbach, 2003; Wayne et al., 2005). Moreover, once a judge-based standard is set, it may be different to the original standard intended by the item/examination writers. This may add further confusion as it is not clear which standard should prevail (Cizek & Bunch, 2007).

The current workshop provides practical and defensible solutions to some of the issues mentioned above: (1) introducing an effective method to determine borderline scores which minimize judges'biases; (2) introducing the use of the Rasch model in setting standards at the item level as well as the whole test level; (3) Introducing a technically simpler method to the Rasch model.

The Rasch Borderline Method (RBM) and related methods discussed in this workshop will demonstrate how defensible Pass/Fail decisions could be made at the item and test levels while minimising the impact of judges' attributes/biases on those decisions (Chang et al., 1996; Wayne et al., 2008). In the workshop we will also introduce results from new simulation studies demonstrating that the accuracy of a Rasch based standard setting method exceeds 70% with approximately 10% false positive and 18% false negative rates.

Thursday April 16, 2015
1:00 PM - 5:00 PM, Seville Ballroom West, Lobby Level, Training Session, TT

An Introduction to Using R for Quantitative Methods

Brian Habing, University of South Carolina and Jessalyn Smith, CTB/McGraw-Hill Education

R has become a favorite of researchers. This half-day course will interactively cover some of the most useful aspects for data analysis and statistical methods. This course is designed for those who are interesting in using R for applying quantitative methods but have no previous experience with R.

This session is designed to introduce the statistical package R so that the attendees will both be able to use R for basic statistical analyses and have an understanding of how it can be used in their own teaching, research, or operational work. This will include guidance on becoming familiar with R, selecting appropriate packages for carrying out more advanced methods, and providing selected custom designed functions to easily produce output in a format similar to that presented in most text books.

Thursday April 16, 2015
1:00 PM - 5:00 PM, Valencia, Lobby Level, Training Session, UU

Understanding Automated Scoring: Theory and Practice

Peter W. Foltz, Pearson, Claudia Leacock, McGraw-Hill Education/CTB, and David Williamson, Educational Testing Service

Automated scoring of constructed responses is increasingly used for formative and summative assessments. The goal of this session is to open up the black box and enable a practitioner's understanding of automated scoring: what it is; what it does and does not do well; and how to use it.

Even as debate rages about whether computers should assign scores to essays, the reality is that automated scoring systems (AES) are part of the assessment landscape. Why? There are clear advantages of automated scoring in terms of reporting time, costs, objectivity, consistency, transparency and feedback. There are also major challenges: Automated scoring systems cannot read or understand an essay in the way that humans, nor are they designed to recognize novelty and they cannot evaluate many aspects of higher-order reasoning. The goal of this workshop is to tease apart what automated scoring systems can do well, what they cannot do well, and what are reasonable short-term expectations for how it can be used based on current research. The workshop is designed to provide practical advice and considerations for practitioners about what automated scoring is, how it works, how it can be applied, and what decisions need to be made in integrating these systems into formative and summative assessments.

Opening the Black Box
We will provide an in-depth orientation to the various methods that are used to create an AES to evaluate both written and spoken assessment responses. Text-processing techniques are deployed to evaluate essays, content-based short-text responses. Speech processing techniques are currently used to evaluate spoken responses. These systems are driven by constructing features and manipulating using statistical tools of NLP/Speech processing and Machine Learning. Other NLP/Speech based systems are used to detect plagiarism, gaming and at-risk essays. We will discuss methods of constructing features and methods of assembling them into scores. We will contrast NLP/Speech tools/orientation to those of psychometric tools, drawing parallels to the measurement concepts the audience might be more familiar with (e.g. polytomous items; conditional dependence problems). Finally, we will distinguish scoring from feedback. Participants will have the opportunity to write (or copy and paste) essays into two operational essay scoring systems to learn about how the systems are used.

Putting Systems into Operational Practice
Designing and implementing an AES system is only the first step. Turning it into an operational system and deploying it is a topic that is usually completely overlooked in research papers on automated scoring. We will describe different models of deployment and their processes. Steps include preprocessing of the textual data, filtering unscorable essays and diverting to hand-scoring, model building, assigning a score and reporting it.

Evaluation of Automated Scoring Systems
The workshop will describe different approaches to the performance of AES systems. The most common approach is evaluation relative to human scores such as exact or adjacent agreement and quadratic weighted kappa. However, flaws in human scoring are well known so that measurement relative to human scores may not be the best evaluation metric. Systems can also be evaluated relative to external criteria and relative to construct definitions. Finally, we will address the most important question -- what is good enough for system deployment?

...continue on page 39

...continued from page 38

Maintaining Systems Over Time
Another critical topic that will be covered is how to maintain an automated scoring system over time. Managing changes to the engines as they are improved has many implications. In addition, as a prompt ages, there are trends/drift over time in both human scores and nature of responses. The workshop will describe how to incorporate these considerations into operational practice.

Open Issues, Future Directions, and General Discussion
The workshop will conclude with a discussion about appropriate applications of automated scoring, issues for its implementation and where the field is going. This section will contrast the current state of tools and knowledge about selected response with approaches for automated scoring of constructed responses. It will examine mechanisms by which the automated scoring can blend some selected response approaches and human constructed response scoring. Finally, the workshop will end with a general discussion about what the field needs, both in terms of NLP/Speech and psychometrics, to strengthen the operational use of automated scoring and provide some predictions for the future..

Thursday April 16, 2015
4:00 PM–7:00 PM, Cordova Room, 5th Floor

NCME Board of Directors Meeting

Members of NCME are invited to attend as observers.

Annual Meeting Program - Friday, April 17, 2015

Friday April 17, 2015
8:15 AM - 10:15 AM, Empire Ballroom, 7th Floor, Coordinated Session, A1

Use of Evidence-Based Standard Setting in PARCC Assessments, Coordinated Session

Session Chair: Walter (Denny) Way, Pearson
Session Discussant: Chad Buckendahl, Alpine Testing

It is recognized that PARCC performance standards must reflect both college and career readiness focus of the assessments and align a system of performance standards across grades and high school courses. This coordinated symposium will discuss research studies defined to support PARCC performance standards to be set in summer 2015.

PARCC Standard Setting: General Approach and Context
Laurie Davis, Pearson & Jason L Meyers, Pearson

The standard-setting process for the PARCC summative assessments will set five performance levels. PARCC will utilize an evidence-based approach to standard setting, involving seven general steps. This paper provides an overview, context, and describes these seven steps to be used in setting performance standards for the PARCC program.

PARCC Benchmarking Study
Katie Larsen McClarty, Jennifer L. Kobrin, Eric Moyer, Sarah Griffin, Kathy Huth, Sharlotte Carey, & Susan Medberry, Pearson

The PARCC benchmarking study gathered external information to estimate the percentage of students college- and career-ready. Through comparisons of performance level descriptors and empirical readiness definitions, the performance level on each external assessment most closely aligned with PARCC's Level 4 was determined and associated student performance data recorded.

PARCC Postsecondary Educators' Judgment Study
Eric L Moyer, Laurie Davis, Julie Miles and Wenyi You, Pearson

The PARCC Postsecondary Educators' Judgment Study used an online standard setting process to collect judgments, from a sample of educators at institutes of higher education, about the minimum performance required to receive the College and Career Readiness Determination. The results of this study inform the Evidence Based Standard Setting (EBSS) process.

International Benchmarking Study—Content Alignment Component
Mary J. Pitoniak, Nancy Glazer, Luis Saldivia, Educational Testing Service

In preparation for statistical linking of PARCC assessments to international assessments (TIMSS, PISA, PIRLS), a content alignment study is being conducted to evaluate the degree to which the blueprints and items on PARCC assessments and those on the international assessments measure the same construct. Preliminary study results will be presented.

Friday April 17, 2015
8:15 AM - 10:15 AM, Exchange, 11th Floor, Paper Session, A2

DIF: Bayesian and Mixed

Session Chair: Bilir Kuzey, Pearson
Session Discussant: Craig Wells, University of Massachusetts Amherst

Differential Item Functioning: Comparison of Traditional and Modern Bayesian Methods
William Muntean, Pearson and Ada Woo, National Council of State Boards of Nursing

To reduce false positives, the current study employs a Bayesian framework and explores Bayes factors incorporating priors on effect size. Requiring more evidence for DIF, this method reduces spurious effects. A comparison is made to traditional approaches (Mantel-Haenszel; likelihood ratio test) and other Bayesian methods (random item mixture model).

A Bayesian Odds Ratio Approach for DIF in Polytomous Items
Oksana Naumenko and Randall Penfield, The University of North Carolina at Greensboro

Traditional and Empirical Bayes (EB) approaches to DIF detection were applied to polytomous items using the Liu-Agresti cumulative common log-odds ratio as the estimator of the DIF effect. A simulation study manipulating DIF form, group size, and impact compared the performance of traditional and EB approaches to DIF parameter recovery.

Using a MIRT Framework to Better Understand Differential Item Functioning
Cindy Walker, University of Wisconsin-Milwaukee and Sakine Gocer Sahin, Hacettepe University

This study's purpose was to determine how distinct the underlying ability distributions for the two studied groups needed to be in order for DIF to occur. Multidimensional data was generated and the underlying ability distributions on the dimensions were manipulated for the reference and focal groups in varying degrees.

Continuous Latent Variables as Covariates in a Mixture Rasch Model
Tugba Karadavut, Allan S. Cohen, and Seock-Ho Kim, University of Georgia, Athens, GA

After fitting a Mixture Rasch Model, adding covariate to the model provides information about the differences between the latent classes. Previous research used manifest grouping variables as covariates which were not informative in DIF studies. In this study, latent continuous covariates are used. Three methods are compared for covariate incorporation.

A Bayesian Approach to Detect Item Parameter Drift Under the NEAT Design
Lihong Yang and Mingcai Zhang, Michigan State University, East Lansing, MI

This paper investigated the effectiveness of the Bayesian approach in detecting Item Parameter Drift (IPD) for common items between two longitudinal English proficiency tests. A Gibbs sampler method was conducted to see how the different NEAT designs, different IPD strengths and sample sizes affect the detection rates.

Mixture-IRT Models With Two Sources of Differential Item Functioning
Erin Strauts and Jessica Flake, University of Connecticut

This study investigates the practicality of using mixture modeling to detect uniform differential item functioning (DIF) in the case of polytomous indicators, binary indicators, and one or two sources of DIF. The ability of mixture item response models (MixIRT) to classify respondents under these conditions is assessed.

Friday April 17, 2015
8:15 AM - 10:15 AM, Grand Ballroom, 7th Floor, Coordinated Session, A3

Various Efforts to Evaluate the Quality of Assessment Programs

Session Chair: Thanos Patelis, NCIEA
Session Discussant: Michael Kane, Educational Testing Service

The purpose of this session is to provide examples of various national and international efforts evaluating the quality of assessment programs. Each presentation will describe the approach, the standards, policies, and procedures used. Specific methodologies and results will be presented. Recommendations of the implications and uses will be provided.

Test Reviewing at the Buros Center for Testing
Kurt F. Geisinger, Buros Center for Testing, the University of Nebraska-Lincoln

The Buros approach to the reviewing of tests is described first. Second, the attempt to standardize reviews is considered. An argument is presented that it is difficult if not impossible to set pre-determined levels of technical characteristics for tests; situations differ too drastically across different venues in education and psychology.

Policies and Procedures for the Independent Evaluation of Assessment Programs
Andrew Wiley, Alpine Testing Solutions

Many testing programs have developed systems for the collection of evidence that demonstrate the validity of their programs, but many others require independent evaluation. Alpine has conducted many of these reviews and this session will focus on the fundamental themes encountered, along with the evidence necessary for each.

Ensuring Assessment Quality: The ETS Internal Audit Process
Cathy Wendler, Educational Testing Service

ETS products and services are periodically evaluated through an internal audit that determines compliance with the ETS Standards for Quality and Fairness. Key elements of the audit process and the role of the standards in ensuring the development and maintenance of fair, valid, and high quality tests will be provided.

Assessment Quality Related to College and Career Readiness Assessments
Erika Hall, Susan Gillmor, Brian Gong, Karin Hess, Scott Marion, and Thanos Patelis, Center for Assessment

The Council of Chief State School Officers has published a set of criteria for evaluating high quality assessments aligned to college- and career-readiness standards. The presentation will describe the process used to operationalize these criteria, and outline key requirements (e.g., participants, evidence). Recommendations for their use will be provided.

Friday April 17, 2015
8:15 AM - 10:15 AM, King Arthur, 3rd Floor, Coordinated Session, A4

Test Score Integrity in the Age of Common-Core Assessments

Session Chair: Ashleigh Crabtree, University of Iowa
Session Discussant: Wayne Camara, ACT, Inc.

Recent high profile cases of serious security violations on state accountability tests have reignited interest in procedures for preventing, detecting and penalizing adult violators. The purpose of this session is to provide guidance to K-12 testing programs in four broad areas: current laws/policies, proactive measures, data forensics and legal alternatives.

Review of Current State Test Security Laws and Policies
Michelle C. Croft, ACT

The presentation examines state test security laws and policies, identifies exemplary statutory and/or regulatory language, and identifies areas of common weaknesses.

Proactive Test Security Measures
Gregory J. Cizek, University of North Carolina -- Chapel Hill

This presentation on proactive measures will argue that test security is most appropriately viewed as a validity concern, describe the challenges of ensuring test score integrity for assessments administered across consortia of states, and provide concrete actions that can be taken to enhance test security and test score integrity.

Post Administration Data Forensics
John Fremer, Caveon

Data Forensics are statistical methods for detecting test fraud. The basic concept is that of looking very carefully at test takers' performances at the group and individual levels. How do "regular" test takers, working on their own, respond? What standards and criteria should be applied to unusualness before taking action?

Legal Alternatives for Confirming and Penalizing Test Security Violations
S.E. Phillips, Assessment Law Consultant

This presentation will discuss the pros and cons of administrative, civil and criminal actions for dealing with test security violations. Differential investigative procedures, types of evidence, standards of proof, penalties and costs will be considered.

Friday April 17, 2015
8:15 AM - 10:15 AM, Renaissance Ballroom, 5th Floor, Invited Session, A5

NCME-NATD Symposium: Implementing the Common Core Assessments at the District and School Levels: Voices from the Field - Overcoming Challenges, Making it Work

Session Chair: Zollie Stevenson, Jr., Howard University
Session Moderator: Elvia Noriega, Richardson Independent School District

This session will focus on how school districts of varying sizes and complexity implemented the Smarter Balanced and Partnership for Assessment of Readiness of College and Careers (PARCC) assessment programs focusing on: technology and infrastructural changes in districts and schools; assessment planning for multiple grades/classrooms; assessment administrations for students with disabilities and English learners; and accountability. Four assessment professionals responsible for implementation of the common core assessments are featured in the symposium:

Didi Swartz, Assessment Director of the Chicago Public Schools will share insights related to the implementation of the PARCC assessments in a school district with 380,000 students while Dale Whittington, Director of Research and Evaluation for the 5,600 student Shaker Heights City (OH) School District will share her experiences implementing PARCC in a small school district.

Bradley McMillen, Assistant Superintendent for the 150,000 student Wake County (NC) Public Schools has overseen the implementation of the Common Core State Standard-linked assessments in his school district in a changing political climate. His experiences will be counter positioned with those of Melanie Stewart, Assessment Director for the 79,000 student Milwaukee Public Schools (SMARTER Balanced).

Elvia Noriega, Executive Director of Accountability and Continuous Improvement for the Richardson Independent School District in Texas will moderate the panel. Texas is a state that is not participating in either the PARCC or SMARTER Balanced assessment consortiums.

Friday April 17, 2015
8:15 AM - 10:15 AM, Seville Ballroom East, Lobby Level, Coordinated Session, A6

Overview: Theories of Action for Performance Assessment in Large Scale Testing Programs

Session Chair: Steve Ferrara, Pearson Research and Innovation Network
Session Discussant: Emily Lai, Pearson

Steve Ferrara, Pearson, Ellen Forte, edCount LLC, and Marty McCall, Smarter Balanced Assessment Consortium

Performance assessment was used in statewide testing programs in the 1990s and is re-emerging in next generation assessments. Claims for its positive influences on teaching and learning were common. This presentation summarizes research on these claims and the relevance to theories of action research needed for next generation assessment programs.

Integrating the Evidence: the NCSC Validity Evaluation in Year 5
Ellen Forte, edCount LLC

NCSC created an alternate assessment for students with significant cognitive disabilities, using an argument-based approach to validation, collecting evidence on each claim from processes and studies. This presentation describes and synthesizes validity evidence from 2014 and 2015 pilots and recent studies that demonstrate how well students engage with assessment materials.

Theories of Action Undergirding WIDA's New English Language Proficiency Assessment
Megan Montee and Dorry Kenyon, Center for Applied Linguistics

This presentation highlights the WIDA Consortium's theory-of-action claims, with particular emphasis on the web-based speaking component. It describes ongoing research that supports these claims, including an analysis of responses to speaking test tasks during field testing and research on the development of new rubrics and scoring materials.

Analysis of Field Test Results for the Smarter Balanced Performance Tasks
Marty McCall, Smarter Balanced

The Smarter Balanced Theory of Action emphasizes deep cross-discipline concepts requiring problem solving, analysis and communication, and complex performances in addition to more traditional tasks based on discrete skills and knowledge. This session explores performance task implementation in pilot and field test results from 2014 and 2015 field test design.

Friday April 17, 2015
8:15 AM - 10:15 AM, Seville Ballroom West, Lobby Level, Paper Session, A7

Item Development

Session Chair: Joni Lakin, Auburn University
Session Discussant: Robert Kirkpatrick, Pearson

Predicting Item Difficulty by Analysis of Language Features
Jeffrey McLeod, Donna Butterbaugh, James Masters, and Emma Schaper, Pearson, Bloomington

Benchmarking the accuracy of a method for predicting approximate item difficulties in lieu of field test data. Input variables are linguistic features of items and content domain codes. Linguistic features are derived using natural language processing methods. Nonparametric regression (Classification and Regression Trees, CART) is used to derive prediction rules.

Modeling Local Item Dependence in Multipart Items Using Item Splitting
Hong Jiao, Robert Lissitz, University of Maryland, College Park; and Enis Dogan, PARCC

This study investigates directional local item dependence in multipart evidence-based selected response items. Different scoring algorithms and calibration methods are compared including two item splitting methods, the standard IRT model, the polytomous IRT model, and the testlet model. Model fit and parameter estimation consistency and accuracy will be reported.

Leveraging Psychometric Isomorphism in Assessment Development
Katie Kunze, Arizona State University and Vandhana Mehta, Bay Area Techworkers for Cisco

Methods for understanding and evaluating psychometric isomorphism will be used to determine the "necessary levels" of similarity for items to be considered psychometrically isomorphic when used in both low stakes and high stakes assessments. These levels could impact operational assessment programs and reduce the potential threat of item exposure.

Item Position Effects in a K-12 Reading Comprehension Assessment
John Denbleyker and Shuqin Tao, NBOME, Rolling Meadows, IL

The present study analyzes item position effects and corresponding impacts on ability estimation. It draws upon a test construction design that had reading passages presented in different positions across multiple forms. Bayesian MCMC methods are used to evaluate credible differences in response probabilities and generate ability estimates.

Item Calibration Using Small Sample Sizes and Short Test Lengths
Alper Sahin, Cankaya University and Duygu Anil, Hacettepe University

Drawing on a large-scale real exam data, this paper focuses on the consequences of calibrating item parameters using small sample sizes and short test lengths while developing language tests through unidimensional IRT. Some practical implications regarding the use of small sample sizes will be presented.

The Impact of Item Compromise on Test Scoring
Jiyoon Park, Zhu Wang, and Lorin Mueller, Federation of State Boards of Physical Therapy

This study evaluated the impact of item compromise on examinees' theta estimates and proposed a way to improve the accuracy of the item calibration process. Results suggested that theta estimates are relatively stable when only noncompromised items are used in the estimation process, regardless of the severity of item compromise.

Friday April 17, 2015
8:15 AM - 10:15 AM, Toledo, 5th Floor, Coordinated Session, A8

Test Batteries Under Sequential Designs: A Technology-Enhanced Examination

Session Chair: Haiyan Lin, ACT, Inc.
Session Discussant: Hua Hua Chang, University of Illinois at Urbana Champaign

Designs of Test Batteries With Adaptive Multistage Testing Models
Wen Zeng, University of Wisconsin at Milwaukee, Haiyan Lin, ACT, Inc., and Cindy M. Walker, University of Wisconsin at Milwaukee

This study proposes two MST-based battery designs. The MST battery (MSTB) design includes three connected MST tests. The hybrid MSTB (MSTBH) design contains the first two tests under MST and the last under CAT. Routing stage assembling approaches of subsequent tests by using ability information from previous test(s) are investigated.

Comparing Between-Item Multidimensional CAT With CAT Batteries
Haiyan Lin, ACT, Inc.

One between-item MCAT (BMCAT) design selects items from a single-content subpool before moving to another single-content subpool. Another design selects items from a pool mixing all contents. BMCAT designs with and without content controls are compared with baseline models of CAT batteries and regular CAT separately administered in three subjects.

Dealing With Not-Reached Items in CAT Batteries
Qing Yi and Meichu Fan, ACT, Inc.

Testing programs often use a test battery to measure several distinct but correlated contents. Researchers have examined the impact of not-reached items on scoring in computerized adaptive testing (CAT) with a single test. This study explores appropriate scoring methods for not-reached items in computerized adaptive test battery (CATB).

Impact of Drifted Items on CAT Batteries
Sien Deng, University of Wisconsin at Madison; Meichu Fan, and Qing Yi, ACT, Inc.

Item parameters may be drifted over time for various reasons. This study investigates item parameter drift (IPD) effects on ability estimations and item pool usage in adaptive test batteries. Simulated IPD conditions include drift type, drift magnitude, drift direction, and percentages of drifted items with two CAT item selection methods.

Friday April 17, 2015
8:15 AM - 10:15 AM, Valencia, Lobby Level, Paper Session, A9

Linking in General

Session Chair: Jade Caines, University of New Hampshire
Session Discussant: Michael Walker, College Board

Examining Alternatives for Linking Rights Scored to Formula Scored Tests
Gautam Puhan and Longjuan Liang, Educational Testing Service

When a testing program moves from formula scoring to rights scoring, it can cause a discontinuity in the score scale unless the rights scored tests forms can be successfully equated to the formula scored test forms. This study examines several alternatives to equate rights scored tests to formula scored tests.

Linking Methods for the Zinnes-Griggs Pairwise Preference IRT Model
Philseok Lee, University of South Florida, Seang-Hwane Joo, University of South Florida and Jacob Seybert, Educational Testing Service

Pairwise preference IRT models pose interesting new questions about linking and DIF detection. Despite their rowing appeal for measuring attitudes and interpersonal skills that predict student performance, little research has been devoted to linking. This study developed and compared four linking methods using a Monte Carlo simulation.

Impact of Linking Designs on Simultaneous Linking
Zhiming Yang, Educational Records Bureau and Shelby Haberman, Educational Testing Service

Simultaneously linking a large number of test forms has been implemented in some testing programs. The impact of linking designs on the performance of this method is examined through a simulation study. Initial results indicate that the dispersed linking design yields better results than the chain linking design.

Linking Composite Scores: Effects of Anchor Test Length and Content
Peng Lin, Neil Dorans, and Jonathan Weeks, Educational Testing Service, Princeton, NJ

This simulation study investigates the effect of content representativeness and the length of anchor test on linking via different methods when the two forms are not parallel in content with data collected from non-equivalent groups.

Friday April 17, 2015
10:35 AM - 12:05 PM, Grand Ballroom, 7th Floor, Invited Plenary Session, B1

The Role of the Measurement Profession in the Renewal of ESEA and Other Federal Education Initiatives
Session Chair: Kristen Huff, Regents Research Fund
Presenter: John King, Senior Adviser Delegated the Duties of Deputy Secretary, USED

Dr. King will address ESEA re-authorization, with particular focus on how the discourse can be better informed by active participation from the measurement community, especially given the latest version of the testing standards. A generous proportion of time will be left for Q&A and discussion.

Friday April 17, 2015
12:25 PM - 1:25 PM, Camelot, 3rd Floor, Electronic Board Session, Paper Session, C1

Electronic Board #1
Impact of College Admissions Measures on Characteristics of Selected Students
Rebecca Zwick, Lei Ye, Steven Isham, and Zhumei Guo, Educational Testing Service

We use nationally representative data on applicants to selective colleges to simulate the effect of competing admissions rules on the demographics (socioeconomic status, gender, ethnicity), academic qualifications (high school GPA, admissions test scores), and college outcomes (graduation rate, grades) of the selected students. These simulations can help inform admissions policy.

Electronic Board #2
Who Does Not Complete the Test?
Marit List, Gabriel Nagy, and Olaf Köller, Leibniz Institute for Science and Mathematics Education, Kiel, Germany

We studied the relationship between ability and tendency to not complete a test for low-stakes mathematics assessments. Two independent large-scale study samples show that larger proportions of not-reached items are related to not only higher levels of ability as measured by the test but also to school grades and IQ-scores.

Electronic Board #3
Using Repeaters' Growth Models to Monitor Test Performance Across Administrations
Youhua Wei, Educational Testing Service, Princeton

For a large-scale and high-stakes testing program, some examinees take the test more than once. This study uses multilevel growth modeling to explore the repeaters' score change patterns and examines its potential for the quality control of test performance across administrations.

Electronic Board #4
Performance Assessment of Physicians' Clinical Skills: A Multivariate Generalizability Analysis
Polina Harik, Brian Clauser, Su Baldwin, Janet Mee and Kimberly Swygert, NBME, Philadelphia

This paper reports on an application of multivariate generalizability theory to assessment of physicians' clinical skills. During the exam, examinees interact with a series of standardized patients and each interaction results in four component scores. The paper focuses on the practical interpretation of the estimates of variance and covariance components.

Electronic Board #5
Investigation of Response Changes in the GRE Revised General Test
Ou Lydia Liu, Brent Bridgeman, Lixiong Gu and Nan Kong, Educational Testing Service, Princeton

This study analyzed examinees' response change pattern and outcomes (i.e., score gain or loss) in the GRE revised General Test, using data from over 8,000 examinees. The analyses yielded findings consistent with prior research that examinees in general benefit from response change, and such benefits increase as examinees ability increases.

Electronic Board #6
Optimal Sample Sizes for Multilevel Latent Class Models
Hsiu-Ting Yu, McGill University, Montreal, Canada

When developing and planning a multilevel study, the sample size at the group and individual levels has to be determined simultaneously for unbiased and accurate estimates. In this paper, simulation studies were conducted to investigate the sample sizes requirement at group and individual level for multilevel latent class models (MLCM).

Electronic Board #7
Impact of Switching between Item Language Versions on Examination Scores
Andre De Champlain, Marguerite Roy, Liane Patsula, and Sirius Qin, Medical Council of Canada

For the current medical licensing examination program, candidates are able to switch freely between English and French versions of an item, on an item by item basis. In this study, we explore the relationship between language switching at the item level and examinee performance at the total test score level.

Electronic Board #8
Modified Generalized Likelihood Ratio and Item Selection Methodology
Samuel Haring, Pearson and Barbara Dodd, University of Texas at Austin

Likelihood ratio-based classification tests including the generalized likelihood ratio (GLR) demonstrate improved test length while maintaining classification accuracy. This paper examines the functionality of a modified GLR procedure which does not incorporate the unnecessary error inherent in the GLR procedure and explore the use of ability-based item selection.

Electronic Board #9
Visualization of Factor Structure of an Instrument Using Multidimensional Scaling
Cody Ding, University of Missouri-St. Louis, St. Louis

In this paper, we proposed to use multidimensional scaling model for visualizing factor structure of an instrument. Traditionally, exploratory factor analysis is a common method for such a purpose. We demonstrate how factor structure of an instrument can be investigated by the visualization method, which can provide an alternative method.

Electronic Board #10
A Two-Step Growth Mixture Modeling With Distributional Changes Over Time
Joseph Nese, University of Oregon and Akihito Kamata, Southern Methodist University

This paper introduces and demonstrates a two-step growth mixture modeling (GMM) analysis when the distributional characteristics of the outcome measure change over time. This paper applies the proposed approach to a data set from kindergarten students on an emergent reading skill (letter sound fluency) across the year.

Electronic Board #11
Investigating the Amount of Systematic and Random Error in Classroom-Level SGPs
Joshua Marland, Craig Wells, and Stephen Sireci, University of Massachusetts Amherst; and Katherine Castellano, Educational Testing Service

Given the increasing use of student growth percentiles (SGPs) to evaluate educational effectiveness, it is important to understand the link between statistical properties and practical implications. This simulation study examined the amount of systematic and random error in classroom-level SGPs as a function of classroom characteristics.

Electronic Board #12
Evaluating Psychometric Properties of Teacher Classroom Observation Ratings
Yi-Chen Chiang and Ginette Delandshere, Indiana University, Bloomington

In this paper we evaluate the psychometric properties of teachers' classroom observation ratings and their associations with student achievement scores and teacher value-added scores to provide insight on the internal and external structure of the construct "teacher effectiveness".

Electronic Board #13
Mapping the Emotion Space With the MGGUM
James Roberts and Jordan Sparks, Georgia Institute of Technology, Atlanta, GA

The MGGUM is a multidimensional IRT model for unfolding responses to tests/questionnaires that presumably follow from a proximity relation. As an illustration, subjects' self-similarity ratings for 24 photos depicting various facial emotions are analyzed to produce a joint map of individual and photograph locations in a multidimensional latent space.

Electronic Board #14
Factor Analyzing Change Scores from the Civic-Mindedness Scale
Monica Erbacher, Kelly Foelber, and S. Jeanne Horst, James Madison University, Harrisonburg, VA

Higher edcation programs often purport to impact students' development of attitudes, such as civic-mindedness. The dimensionality of students' (N = 602) pre- to post-test change on the Civic-Mindedness Scale (Hatcher, 2008) was explored. Findings suggest change dimensionality differs from single occasion dimensionality. Implications for understanding student development are discussed.

Electronic Board #15
Predicting Item Parameters Using Regression Trees
Jeffrey Steedle and Steve Ferrara, Pearson, Austin, TX

Data from national achievement test items were analyzed using regression tree analysis with conditional random forests to determine what item features predicted difficulty and discrimination. The statistical models accounted for 2% to 46% of the variance in item parameters; item type and standard alignment were the most important predictors.

Electronic Board #16
Predicting Students' Writing Performance on NAEP from Assessment Variations
Ya Mo and Gary Troia, Michigan State University, East Lansing, MI

The research examines the relationship between students' NAEP performances and the amount of difference between state and NAEP direct writing assessments through HLM and finds that students' preparedness for the tasks, namely the similarity between the assessments of their home states and the NAEP, plays a role in students' performance.

Electronic Board #17
Bayesian Networks to Model Science Inquiry Skills in NAEP ICTs
Johnny Lin, Margarita Olivera-Aguilar, and Yue Jia, Educational Testing Service

The NAEP 2009 interactive computer tasks pose psychometric challenges for modeling student observables. A proposed solution is the Bayesian Network (BN). This paper demonstrates how BNs can model complex conditional relationships among observables and create tailored proficiency profiles of student scientific inquiry skills.

Electronic Board #18
Performance Benchmarks for GDCM-MC Option-Based Modeling
Louis DiBello, William Stout, Learning Sciences Research Institute at University of Illinois at Chicago; and Robert Henson, UNCG

For best practice applications of the GDCM-MC model to diagnostic assessment settings, multiple choice option-based scoring is investigated through a comprehensive series of simulation studies based on real data analyses. Definitive measures of model-data fit, discriminability, and classification accuracy are derived and applied to Q matrix specification and test development.

Electronic Board #19
Bayesian Multilevel Multidimensional Item Response Theory Model
Ken Fujimoto, Loyola University Chicago, Chicago, IL

The multilevel multidimensional item response theory (MMIRT) model is generally expressed within a Bayesian framework. Doing so relaxes some of the restrictions that have to be placed on model parameters under a frequentist framework. The advantages of the Bayesian MMIRT are shown through the analysis of real rating data.

Electronic Board #20
Classification Issue Using the Diagnostic Classification Models
Yuehmei Chien, Pearson, Ning Yan, Independent Consultant, and John Behrens, Pearson

Setting up an indifference region (such as [0.4, 0.6]) to obtain classification decisions—non-mastery, indifference, and mastery—on each attribute might not produce results as expected when only few students fall into the indifference region. This paradox will be described and presented through simulation.

Electronic Board #21
Confidence Intervals for Mastery Probabilities of Attributes
Chingwei David Shin, Yuehmei Chien, Pearson, Ning Yan, Independent Consultant

To provide a confidence interval around the mastery probability of an attribute, reliability is an option but not ideal. We first conducted a comparison study to show that the existing attribute reliability performed differently. Then, a bootstrap approach was proposed to derive the confidence interval, which showed promising results.

Electronic Board #22
Evaluating the Implementation of Diagnostic Classification for CAT
Hui Deng, the College Board, Yuehmei Chien, Chingwei David Shin, Pearson

The potential benefit of using DCM for a state-mandated CAT assessment was explored. Data from a large sample were used to calibrate both the IRT and DCM parameters. The traditional IRT-CAT was compared to the IRT-DCM-CAT, which took into account both theta and attribute-vector estimates in item selection.

Electronic Board #23
The Use of Support Vector Machine in Cognitive Diagnostic Assessments
Cheng Liu, Ying Cheng, University of Notre Dame

We propose to use the support vector machine (SVM) to make classification decisions on each attribute given a training dataset. By using SVM we evade fitting and calibrating a CDM, and estimating a latent class. Instead, it becomes a quadratic optimization problem in hyper dimensional space. Classification parameters determining the fail or pass boundary can be obtained from the training dataset before operational testing. No large calibration sample is required and no O(2k) computational time is needed.

Electronic Board #24

Investigating Practical Utility of Meta Analysis for DIF Investigation

Jin Koo, American Nurses Credentialing Center and Insu Paek, Florida State University

Practical utility of the meta-analysis technique for the investigation of DIF (meta-analytic method for DIF detection) is evaluated and compared with the Mantel-Haenszel method and the Breslow-Day test under various conditions. Our preliminary analysis showed that the meta-analytic method performed as good as the MH procedure for detecting uniform DIF.

Electronic Board #25

Effects of Scoring Designs on Rater Precision and Classification

Yoon Soo Park, University of Illinois at Chicago; Mi Hwa Kim, Seoul National University, and Kuan Xing, University of Illinois at Chicago

In many large-scale assessments, subsets of examinee responses from constructed-response items are rescored to ensure inter-rater reliability. This study examines the effect of scoring designs by varying characteristics of the primary rater using a rater model (latent class signal detection theory model). Implications for rater precision and classification are discussed.

Friday April 17, 2015
12:25 PM - 1:55 PM, King Arthur, 3rd Floor, Invited Session, C2

Spencer Foundation: From Funded to Unfunded: What Makes the Difference

Amy Dray and Amy Proger, Program Officers, The Spencer Foundation

What makes the difference between funded and unfunded grant proposals in education? In this session, program officers from the Spencer Foundation will identify the key ingredients that distinguish winning proposals from proposals that are unsuccessful. They will also talk about the art of reshaping unsuccessful proposals to be competitive. Among the questions to be addressed: What are the defining qualities of proposals that receive financial support? How do early-career scholars garner funding? What are the limitations of proposals that fall short of their funding goals? The session will include a question and answer period.

Friday April 17, 2015
12:25 PM - 1:55 PM, Renaissance Ballroom, 5th Floor, Coordinated Session, C3

Measuring Students' Proficiency on the Next Generation Science Standards

Session Chair: Judith Koenig, National Academy of Science/National Research Council
Session Discussant: Joanna Gorin, Educational Testing Service and Lauress Wise, HumRRO

Presenters: James Pellegrino, University of Illinois at Chicago, Mark Wilson, University of California, Berkeley, and Joan Herman, University of California, Los Angeles

The 2012 Framework for K-12 Science Education and the Next Generation Science Standards call for a new approach to science education that focuses on the integration of core ideas, practices, and crosscutting concepts. New approaches to assessments will be needed. This session discusses these approaches and explores associated psychometric challenges.

Friday April 17, 2015
12:25 PM - 1:55 PM, Toledo, 5th Floor, Coordinated Session, C4

Automated Scoring, Item Generation and Mixed-Format Adaptive Testing

Automated Scoring and Adaptive Test Designs
Isaac I. Bejar, Educational Testing Service, Princeton, NJ

This paper will discuss automated scoring in the context of mixed-format designs and will briefly describe the anatomy of automated scoring engines.

An Adaptive Mathematics Assessment with On-The-Fly Item Generation
Meirav Attali and Yigal Attali, Princeton, NJ

This presentation illustrates the simultaneous integration of automated scoring and item generation in the context of a middle-school mathematics test and presents empirical results about the psychometric feasibility of the approach.

Automatic Generation and Scoring of Mathematics Items
James Fife, Princeton, NJ

This presentation illustrates the joint use of item generation and automated scoring by means of constructed-response item models that, in addition to aiming to generate items of comparable difficulty, also generated the scoring key. Results will be presented on how successfully psychometrically comparable items were generated and scored.

Integrating Difficulty Modeling and Item Generation for Vocabulary Items
Paul Deane, Princeton, NJ

Item generation methods can also be used to generate items that vary in difficulty, as illustrated by this presentation. To do so, however, a difficulty model is needed. The presentation illustrates the formulation of a difficulty model for vocabulary items based on natural language processing attributes of the items. Those attributes are, in turn, the basis for the generation of additional items.

Combining Automated Scoring Constructed Responses and Computerized Adaptive Testing
Qiwei (Britt) He, Princeton, NJ

This presentation illustrates a mental health application with a test design consisting of an essay as the "first stage" followed by a multiple-choice adaptive test. Results are presented that suggest that measurement accuracy was improved by incorporating the essay as a routing test.

Friday April 17, 2015
12:25 PM - 1:55 PM, Seville Ballroom East, Lobby Level, Coordinated Session, C5

Methodological Developments in International Large-Scale Assessments

Dubravka Svetina, Indiana University, Matthias von Davier, Educational Testing Service, David Kaplan, University of Wisconsin-Madison and Eugene Gonzalez, Educational Testing Service

In this coordinated session, we present several new quantitative methods relevant for international assessments, such as PISA and TIMSS. In particular, we present papers on cross-cultural equivalence fit measures, causal inference using a Bayesian framework, and a 'big-data' approach to linking multiple study cycles.

Friday, April 17, 2015
12:25 PM-1:55 PM, Exchange, 11th Floor, Invited Session, C6

Handbook of Test Development (2nd Ed): Major Advances and Implications for Test Developers

Session Chair: Suzanne Lane, University of Pittsburgh
Session Discussant: Suzanne Lane, University of Pittsburgh

This session introduces the 2nd edition of the Handbook of Test Development. Similar to the original Handbook, this edition documents sound testing practices in a way that is useful to both test developers and researchers studying issues that affect test development.

Foundations in Test Development and Advances in Delineating the Content and Skills to Assess
Suzanne Lane, University of Pittsburgh

The foundations of test development are introduced, including evidence-centered design, validity and fairness issues in testing. Advances in methods for specifying the content to be measured and what is measured will also be discussed.

Advances in Item Development: Web-Based Item Development to Automated Item Generation
Mark Raymond, National Board of Medical Examiners

Advances in item development and scoring will be addressed, including web-based item development, performance tasks, computerized innovative item formats, automated item generation, and automated scoring.

Advances in Test Design Strategies, Reporting, Documentation, and Evaluation
Thomas Haladyna, Arizona State University

Current thinking in developing test forms, vertical scaling, designing computer adaptive tests and automated test assembly as well as advances in test reporting, documentation and evaluation will be discussed.

Use of the Revised 2014 Standards for Educational and Psychological Testing in Test Development
Lauress Wise, HumRRO

The implications of the revised 2014 *Standards* for test design and development are discussed, with a focus on the standards and principles in the test development, validity, reliability and fairness chapters.

Friday, April 17, 2015
12:25 PM-1:55 PM, Empire Ballrooom, 7th Floor, Invited Session, C7

Model Fit and Scoring Invariance Across Multiple Populations

Assessing Item Fit When Items are Reused
Frederic Robin and Sooyeon Kim, Educational Testing Service

After a brief introduction to this symposium, this presentation will first discuss some of the challenges associated with reusing items. It will then focus on the potential of residual analyses to detect item misfit when it reaches a level at which it may affect test scores.

Potential of Time Series and Statistical Process Control Analyses to Detect Item Drift
Hongwen Guo and Frederic Robin, Educational Testing Service

When items are reused over time, the early detection of item drift becomes an important issue. This session will focus on a new item difficulty index that forms a time series and is comparable across administrations. The potential use of time series analyses to detect misfit under various conditions will be discussed.

An Empirical Investigation of the Potential Impact of Item Misfit on Test Scores
Sooyeon Kim and Frederic Robin, Educational Testing Service

Abstract: This study examined the potential impact of item misfit on the reported scores of an admission test from the subpopulation invariance perspective. We compared new conversions derived from empirical data to the conversions derived from the original item parameters to determine whether subpopulation invariance was achieved at the score level.

Alternative IRT Models for Tests with Diverse Test Taking Population
Deirdre Kerr and Frederic Robin, Educational Testing Service

Despite all the challenges it creates, the use of tests across multiple populations is becoming common practice. In particular, with the requirement that one common scoring model be used, the choice can be difficult. In this presentation we will report on the extent to which unidimensional and multidimensional IRT models may fit all of the subpopulations.

Friday, April 17, 2015
12:25 PM-1:55 PM, Grand Ballroom, 7th Floor, Invited Session, C8

Quality Focus: Experiences from a Number of Assessment Programs

Session Chair: Judith Monsaas, University of North Georgia
Session Discussant: Henry Braun, Boston College

Maintaining Fidelity to the Construct Through Stealth ECO
Kristen Huff, Jason Schweid, David Abel, and Peter Coe, Regents Research Fund

At the outset of assessment design, there needs to be reasonable confidence — and evidence — that the intended construct is measured with fidelity; otherwise, the entire validity argument is compromised. The design documents undergirding the New York State Common Core assessments will be discussed in this presentation.

Computer-Based Assessments in the Sunflower State
Marianne Perie, Center for Educational Testing & Evaluation

Kansas is administrating new ELA and math assessments aligned with College and Career Ready Standards. Work on these assessments began in 2012 as a transition to prepare students for Smarter Balanced assessments. In December 2013, the State Board voted to leave the consortium and develop state-specific assessments. The design, process, and planned analyses will be described.

Next-Generation Criteria for Evaluating Assessment Quality
Brian Gong, Center for Assessment

What criteria and procedures may be useful for evaluating the quality of assessment programs, including considerations such as alignment of CAT "test forms," complex content specifications (e.g., text complexity; mathematical/science skills/practices crossed with content knowledge), ALDs viewed as claims, and evaluation of consequences/uses as well as score interpretation? The Center for Assessment has worked on several projects that help provide answers to these questions, including a project to operationalize the CCSSO Criteria for Procuring and Evaluating High Quality Assessments

Assessment Quality Considerations for SBAC
Joe Willhoft, SBAC

The quality issues attended to by Smarter Balanced assessments will be illustrated through a discussion of several key activities over the course of the four-year development period, including: creation of "Content Specifications" and test blueprints, quality control of item development, and an alignment model suitable for adaptive testing.

Friday April 17, 2015
12:25 PM – 1:55 PM, Valencia, Lobby Level, C9

Peer Review of Peer Review

Session Chair: Ellen Forte, Ed Count, Inc.
Session Discussant: Scott Marion, Center for Assessment

Under the two most recent reauthorizations of ESEA, each US state has been required to submit packages of evidence to the U.S. Department of Education for peer review. This session will examine current plans for peer review and offer recommendations based on history and assessment quality reviews conducted elsewhere.

Peer Review in Policy and Practice
Ellen Forte, edCount, LLC

The paper offers a brief history of US peer review and suggests a new vision based on the recently revised Standards (AERA/APA/NCME, 2014) and key developments in validity evaluation and item- and test-development practices. The author offers specific recommendations for a more effective process that is formative and collaborative.

Accreditation of Credentialing Programs: An Analogous Model for Educational Assessment?
Chad Buckendahl, Alpine Testing Solutions

Peer review processes are common for accrediting bodies; this paper describes how an analogous model for educational assessment programs could be developed. The author provides an illustration of how an accreditation model currently used in the credentialing sector could be adapted to serve the needs of educational assessment programs.

Evaluation of Quality of Educational Assessment in the Netherlands
Anton Beguin, NCIEA

Tests that schools in the Netherlands use to evaluate student proficiency are accredited based on a quality check using a set of standards by COTAN (a division of the Dutch association of Psychologists). This paper provides a comparison of the COTAN evaluation system and the new Standards (AERA/APA/NCME, 2014).

An Approach for Evaluating Assessment Quality Involving Peer Review
Thanos Patelis, NCIEA

This paper outlines the role of peer review in the scientific process, referring to its historic roots and summarizing efforts to study its characteristics. The author suggests a methodology that blends existing audits of assessment quality with standards (AERA, APA, NCME, 1999) and best practice criteria (CCSSO & ATP, 2010).

Friday, April 17, 2015
12:25 PM - 1:55 PM, St. Clair, Upper 5th Floor, Paper Session, C10

Considerations for Measuring Item Difficulty

Session Chair: Anna Topczewski, Pearson
Session Discussant: Terry Ackerman, University of North Carolina Greensboro

The Impact of Media Enhancements on Item Difficulty
Amanda Soto, Carol Morrison, and Stephanie Woodward, National Board of Medical Examiners, Philadelphia

Maintaining item statistics that span forms or administration years is particularly important when common items are the basis for equating. This paper investigates the intersection of media and item revision by examining the introduction of pan and zoom capabilities, and the resulting impact on the items' suitability for equating.

Bias Reduction of Gaussian Kernel Smoothed Empirical Item Response Curves
Samuel Livingston, Hongwen Guo, Gautam Puhan, Educational Testing Services; and Allison Ames, University of North Carolina Greensboro

Smoothing of item response curves can introduce bias into the estimated conditional p-values. This research presents a modified smoothing procedure which reduces the bias introduced, but maintains the smooth nature of the curves which test developers rely on to help meet specifications when assembling new test forms.

Exploring the Relation Between Mathematical Content, Cognitive Complexity, and Item Difficulty
Deni Basaraba, Leanne Ketterlin-Geller and Pooja Shivraj, Southern Methodist University, Dallas, TX

Calls abound for mathematics instruction and assessment to promote deeper levels of mathematical thinking. In this study, we examine whether the difficulty of mathematics items is a function of cognitive complexity, the sophistication of knowledge and skills associated with the mathematical content, and whether these relationships are consistent across grades.

Describing Speededness for Computer-Based Tests Using the Time Sensitivity Index
Shu-chuan Kao, Pearson and J. Carl Setzer, GED Testing Service

The time sensitivity index is proposed to reflect the impact of speededness for computer-based tests calibrated by using the Rasch model. This index reflects the difference between item difficulty estimates calibrated from two item latency groups. The method has potential application in form assembling and test design validation.

Friday April 17, 2015
2:15 PM - 3:45 PM, Camelot, 3rd Floor, Electronic Board Session:
GSIC Graduate Student Poster Session, D1

Graduate Student Issues Committee

Lisa Beymer, Chair
Laine Bradshaw, Jeremy Brown, Laurie Davis, Jerusha Gerstner, Jason Herron, Evelyn Johnson, David King, Ray Reichenberg, and Ting Wang

Electronic Board #1
The Effect of Response Style Adjustments on Measures of Variability
Bruce W. Austin, Brian French, and Olusola Adesope, Washington State University, Pullman, WA

Adjustments for response styles from culture or response fatigue are implemented on teacher and student responses from the PISA and TALIS surveys. Aquiescence, disacquiescence, and extreme response styles are investigated. Adjustments to measures of variability for student anxiety, self-efficacy, and teacher-student relations are compared for their effects on teacher outcomes.

Electronic Board #2
Detecting Cluster Bias in a Multilevel Item Response Model
Woo-yeol Lee and Sun-Joo Cho, Vanderbilt University, Nashville, TN

Cluster bias can be investigated by testing whether the within-level item discriminations are equal to the between-level item discriminations in a multilevel item response model. This study evaluates cluster bias detection methods and presents the impact of ignoring cluster bias on item parameter estimates and person scores.

Electronic Board #3
A Comparison of Item Parameter Recovery Using R-Packages and flexMIRT
Taeyoung Kim and Insu Paek, Florida State University, Tallahassee, FL

This study conducts a comparison study via simulations among a well-known commercial IRT program and five R-IRT packages for the 2PL model with respect to how precise these programs can recover standard error of item parameter estimates as well as item parameters themselves

Electronic Board #4
Classroom Level Influence in Multilevel IRT School Effectiveness Research
Katherine Marino and Pui-Wa Lei, Pennsylvania State University, Bernardsville, NJ

Correct modeling of educational data is critical in identifying influential characteristics of successful schools. While effects of omitting levels of nested data on GLM have been documented, the current study applies two- and three-level IRT models to three-level data to determine consequences of disregarding nesting on IRT item data.

Electronic Board #5
Evaluating the Use of Principal Component Analysis in Q-Matrix Construction
Olasumbo Oluwalana, Chia-Yi Chiu, and Immanuel Williams, Rutgers University, New Brunswick NJ,

A primary purpose of cognitive diagnostic models is to classify examinees based on the presence of attributes or latent skills required to correctly answer test items specified in a Q-matrix. Principal component analysis can augment the construction of the Q-matrix by identifying components on which items load.

Electronic Board #6
Bayesian versus Frequentist Multiple Regression
Ryan Derickson, VHA National Center for Organizational Development and Lihshing Wang, University of Cincinnati

Bayesian and Frequentist methods may produce different estimates under various conditions. This study simulates data with varying sample sizes and degrees of skew to evaluate points of agreement and disagreement between Bayesian and Frequentist multiple regression.

Electronic Board #7
Variable Selection for Rasch Model Using Multiple Bayesian Elastic Net
Ping Yang, University of Missouri and Guohui Wu, SAS

In the presence of highly correlated covariates, the multiple Bayesian elastic net (MBEN) is preferred due to its better performance than many other variable selection methods. We propose variable selection for one parameter Rasch model using the MBEN prior and develop well customized sampling algorithm to achieve efficient implementation.

Electronic Board #8
Exploring and Comparing High Stakes Writing Test Prompts Content Structure
Abdolvahab Khademi, University of Massachusetts Amherst, Amherst, MA

A principal argument in test validation is that a test should assess the intended construct independent of other construct-irrelevant factors. The present study attempts to explore and compare TOEFL and IELTS writing prompts content structures for likely nonlinguistic dimensions such as cognitive complexity, task specificity, topical familiarity or culture-specific factors.

Electronic Board #9
Pretest Item Selection for Online Calibration in Multidimensional Computerized Testing
Rui Guo and Huahua Chang, University of Illinois at Urbana Champaign, Union City, CA

Item parameter calibration is important in IRT based tests. Online calibration dynamically selects pretest items during the operational test, which improves the calibration efficiency. This study compares different pretest item selection methods for online calibration under multidimensional computerized adaptive testing. Results show that item-centered selection methods outperformed examinee-centered methods.

Electronic Board #10
Multilevel Modeling of Predictors of Mathematics Achievement in Ghana
Paul Butakor, University of Alberta, Edmonton, Canada

One of the methodological errors in the school effectiveness research literature is the ecological problem in the analysis of nested data. This study corrects this error by using Hierarchical Linear Modeling (HLM) and the 2007 TIMSS data to model factors accounting for the low performance of Ghana's students in mathematics.

Electronic Board #11
Assessing Alternative Item Fit Indices Under Polytomous IRT Models
Yulim Kang and Guemin Lee, Yonsei University, Seoul, Republic of Korea

The current study is conducted to investigate the performance of four alternative item fit indices, developed by Orlando & Thissen (2000) and Stone (2003), compared to traditional item fit statistic G2 under the GRM and the GPCM. Simulation conditions are as follows; test lengths, response categories and ability distributions.

Electronic Board #12
A Comparison of Three Calibration Methods in Vertical Scaling
Juyeon Lee, Guemin Lee, and Sang-Jin Kang, Yonsei University, Seoul, Republic of Korea

This study aims to investigate how FIC performs and clarifies the relative appropriateness of FIC in vertical scaling through the comparisons with separate and concurrent calibration. In addition, three additional factors were considered: type of proficiency score distribution, sample size, and number of common items.

Electronic Board #13
IRT Scale Transformation Using Different Types of Anchor Tests
Sujin Yang, Guemin Lee, and Sangjin Kang, Yonsei University, Seoul, Republic of Korea

The purpose of this study was to compare results of IRT scale transformation using different types of anchor tests considering the effects of scale transformation methods and sample sizes at the same time.

Electronic Board #14
Dimensionality Assessment of Unfolding Models
Elizabeth J. Williams and James Roberts, Georgia Institute of Technology, Atlanta, GA

A simulation study will be conducted to investigate the performance of a PCA for dimensionality assessment with unfolding data generated by the Multidimensional Generalized Graded Unfolding Model (MGGUM). The expected results are that the PCA will generally identify r+1 dominant dimensions, where r is the number of true dimensions.

Electronic Board #15
Classroom Assessment Scoring System in Diverse Settings: Confirmatory Factor Analysis
Tianna Floyd, Jacqueline Towson, Nicole Terry Patton, and Gary Bingham, Georgia State University, Atlanta, GA

This poster explores the underlying factor structure of the Pre-K Classroom Assessment Scoring System (CLASS) to identify its validity and reliability in urban settings, which are both culturally and linguistically diverse. There exist potentially important policy implications in the accuracy of measurement of the CLASS across diverse early childhood settings.

Electronic Board #16
Linking With Planned Missing Data: Concurrent Calibration With Multiple Imputation
Min Sung Kim and William Skorupski, University of Kansas, Lawrence, KS

A new linking method, concurrent calibration with multiple imputation (CCMI), is introduced and evaluated using a Monte Carlo simulation study. MSE and BIAS pattern of IRT parameters are examined with respect to three crossed factors: linking methods, population distributions, and anchor test length.

Electronic Board #17
Classification Accuracy of Mixture IRT and Cognitive Diagnostic Models
Diego Luna Bazaldua and Young-Sun Lee, Teachers College, Columbia University, New York City, NY

Monte Carlo simulations of multidimensional data resembling real-test situations will be generated under different simulation conditions to examine classification accuracy of examinees and item parameters obtained using mixture item response theory models and cognitive diagnostic models (CDMs).

Electronic Board #18
The Impact of Item Parameter Drift on an Adaptive Test
Beyza Aksu Dunya, University of Illinois at Chicago, Chicago, IL

This study employed a series of simulations to examine impact of item parameter drift on person ability estimates and pass/fail decisions when direction and magnitude of drift change. Based on the simulation results, although person parameter estimates were impacted by the extreme conditions of drift, decision consistency remained high.

Electronic Board #19
Application of Mixture IRT to Online Social Networking Behaviors
Kuan Xing, Yoon Soo Park, and Theresa Thorkildsen, University of Illinois at Chicago, Chicago, IL

This study proposes a mixture IRT application to detect patterns of online social networking activity on Facebook. Mixture IRT models were fit based on empirical data from student responses to social interactions, and simulation studies were conducted to examine bias in recovery of parameter estimates and classification.

Electronic Board #20
Using Classification Tree Models and Bagging to Determine Course Placement
Chansoon Lee and James Wollack, University of Wisconsin, Madison, WI

The purpose of this study is to use classification tree models and bagging to find reliable and valid cut-off scores for placement into college mathematics courses. The effectiveness of tree models, which are new methodologies for determining placement, will also be compared to the commonly used approach of logistic regression.

Electronic Board #21
Outcome Measurement in Educational Evaluation: A Simulation of Assessment-Intervention Alignment
Joshua Sussman, University of California Berkeley, Berkeley, CA

Investigators who evaluate the efficacy of educational interventions must measure achievement with assessments sensitive to intended effects of interventions. This simulation study conceptualizes sensitivity as assessment-intervention alignment. Item response models generate assessment data and model the influence of alignment on statistical power and the validity of a summative program evaluation.

Electronic Board #22
A Hybrid Model Approach to External Multidimensional Unfolding
Matthew Barrett and James Roberts, The Georgia Institute of Technology, Atlanta, GA

This study uses external multidimensional unfolding methods to determine response processes used by individuals making physical attraction judgments. Additionally a hybrid approach is developed that better captures the nature of attraction ratings by simultaneously implementing either a vector or unfolding model for different dimensions of the latent space.

Electronic Board #23
Comparison of IRTPRO, Mplus and WinBUGS: Bayesian Item Parameter Recovery
Nedim Yel, Arizona State University, Tempe, AZ

This simulation study compares the item parameter recovery performance of IRTPRO, Mplus and WinBUGS using Bayesian estimation. The factors manipulated include; sample size, test length, ability, and prior distributions.

Electronic Board #24

An Application of Generalizability Theory to Standard Setting

Seohong Park and Won-Chan Lee, University of Iowa, Iowa City

The generalizability theory is applied to a standard setting of a large scale assessment. The study design is i x (r:g) x o. In the D-study, the number of roundings is three or four, and the number of groups and judges is 2~5, respectfully. Total 32 D-study conditions were conducted.

Electronic Board #25

Comparing Student Growth and Scale Score Across Grade Configurations

Hongyu Diao and Lisa Keller, University of Massachusetts Amherst, Amherst, MA

For the past decades, there has been a movement away from the middle school configuration towards a K-8 configuration across states. The present research uses two-way ANOVA to analyze how grade structure and district location impact student academic performance across different grades.

Friday April 17, 2015
2:15 PM - 3:45 PM, Empire Ballroom, 7th Floor, Coordinated Session, D2

Applications of Model-Based Rater Monitoring Procedures

Session Chair: Brian Patterson, Pearson
Session Discussant: Lawrence T. DeCarlo

An overview paper provides a framework for indices of rating quality and the three remaining papers take three different approaches to providing meaningful feedback to raters and the testing programs that they support; specifically a non-parametric approach, one grounded in item response theory, and another based on signal detection theory.

Exploring the Quality of Expert Ratings Using Mokken Scale Analysis
Stefanie A. Wind, Georgia Institute of Technology and George Engelhard, Jr., The University of Georgia

Distinguishing Several Rater Effects With the Rasch Model
Tian Song and Edward W. Wolfe, Pearson

Incorporating Expert Ratings into Rater Monitoring via Signal Detection Theory
Brian F. Patterson, Pearson; Stefanie A. Wind, Georgia Institute of Technology; George Engelhard, Jr., The University of Georgia

Friday April 17, 2015
2:15 PM - 3:45 PM, Exchange, 11th Floor, Coordinated Session, D3

Assessment for Innovative Learning Technology: Modeling Sources of Dependence

Session Chair: Peter Halpin, New York University
Session Discussant: Robert Mislevy, Educational Testing Service

Learning technology offers rich data for developing novel educational assessments. However, these data typically involve sources of statistical dependence that violate the assumptions of conventional psychometric models. The focus of this symposium is to provide models of student ability that address these new sources of dependence.

Using Simulation to Explore Gender and Cultural Differences During Collaboration
Jessica Andrews and Alina A. von Davier, Educational Testing Service

This paper describes the use of a simulation-based collaborative problem-solving task to examine gender and cultural differences in cognitive and collaborative skills. Analyses explore how patterns of interaction may differ according to gender and cultural background, and how particular patterns are related to performance outcomes.

Modeling the Effects of Collaboration on Mathematics Performance
Peter Halpin, New York University and Yoav Bergner, Educational Testing Service

We describe a likelihood ratio test for the effect of collaboration on academic performance, and outline a pre-test / post-test design in which the effect has a clear causal interpretation. The results are illustrated with data collected using the Edx platform and questions from the NAEP grade 12 math assessment.

Inferring Student Ability Based on Within-Game Actions
Michelle M. Lamar and Malcolm Bauer, Educational Testing Service

We show how a cognitive model for sequential decision-making, the Markov decision process, can be used as a measurement model given the complex process data which are available from many educational games. Recovery of student ability is demonstrated through simulation and the model is applied to data from SimCityEDU.

A Bayesian Framework for Adaptive Learning In Educational Games
Josine Verhagen, Unversiteit van Amsterdam

This paper describes a Bayesian framework for adaptive learning in which learner information and prior game play results are used to inform adaptive games for assessment and learning. The initial findings from an adaptive game related to shape and pattern recognition in preschoolers are presented to illustrate the framework.

Friday April 17, 2015
2:15 PM - 3:45 PM, Grand Ballroom, 7th Floor, Invited Session, D4

Advances in Test Score Reporting

Session Chair: Ronald K. Hambleton, University of Massachusetts Amherst
Session Discussant: Ronald K. Hambleton, University of Massachusetts Amherst

For much of the history of educational testing, test score reporting has been given very limited attention. It is rare to find articles in the literature or even technical manuals prior to 1995 that have investigated best practices in score reporting. Today the situation has changed tremendously and the topic is one of the themes of this year's annual NCME meeting. The presenters are among the most productive researchers and the first, will address subtest score reporting. Every user of test scores it seems is asking for diagnostic information in the form of subtest scores. This first presentation will address the many advances on this topic. In the second presentation, assessment results communication will be addressed from many perspectives including psychometric, legal, political, semiotic, user-acceptance and production or engineering considerations and the complex trade-offs among them. All too often these many dimensions relevant in reporting are not considered and less than satisfactory reporting is the result. In the third presentation, focus will center on promoting valid score interpretations through effective reporting including the use of visualizations and graphics. References too to the new AERA, APA, and NCME Test Standards and their implications for score reporting will be addressed.

Diagnostic Score Reporting: A Review of the Status Quo
Sandip Sinharay, Pacific Metrics, Shelby Haberman, Educational Testing Service, Chun Wang, University of Minnesota, and Gautam Puhan, Educational Testing Service

High Dimensional Projections of Assessment Results Communication
John Behrens, Pearson

Reporting With Purpose: Current Approaches to Promoting Test Score Meaning
April L. Zenisky, University of Massachusetts Amherst

Friday April 17, 2015
2:15 PM - 3:45 PM, King Arthur, 3rd Floor, Coordinated Session, D5

Improving Test Security for State Assessment Programs: Lessons Learned

Session Chair: Marianne Perie, Center for Educational Testing and Evaluation
Session Discussant: Barbara Plake, University of Nebraska-Lincoln, Emeritus

John Olson, Olson Educational Measurement & Assessment Services, John Fremer, Caveon Test Security, William Skorupski, University of Kansas, and Barbara Plake, University of Nebraska-Lincoln, Emeritus

Cheating and test piracy (stealing of test forms or items) pose major threats to the validity of test score interpretation and the credibility of large-scale assessment programs. This session focuses on resources/methods to assist states in improving test security and best practices for the prevention and detection of cheating.

Lessons Learned in Improving Test Security for States: An Overview/Summary
John F. Olson, Olson Educational Measurement & Assessment Services

The purpose of the test security project—to produce a TILSA/multi-state contribution that brings together the best practices of assessment staff and testing organizations and focuses on exemplary prevention and detection procedures to minimize testing irregularities and stop cheating—will be described by the presenter, who worked closely with the TILSA Test Security (TS) Workgroup to direct the projects, co-author the Guidebook and Lessons Learned reports, and ensure the information in the reports was practical and useful to assessment staff.

Lessons Learned in Improving Test Security for States: Findings and Recommendations
John Fremer, Caveon Test Security

The TILSA Guidebook has proved to be a very welcome resource to state and district staff, and the new Lessons Learned report is an excellent addendum to it that provides practical advice and additional guidance. Parts of the Guidebook have been widely used in state workshops and training materials. In this session the effective strategies and practices that states are using to prevent cheating will be presented, as well as their use of data forensics results to detect irregularities and possible improprieties.

Cheating Statistical Methods for Evaluating Test Security: What States are (or Should Be) Doing
William Skorupsky, University of Kansas

The purpose of this research is to identify methods for evaluating group-level aberrance as potential evidence of cheating. These methods will focus on the detection of (1) unusual score gains, (2) erasure or answer-changing behavior, and (3) changes in school demographics (as evidence of purposely not testing certain demographic groups to improve school-level outcomes). The paper will provide the technical details behind these methods, demonstrate their use with data from several anonymous state testing programs, and discuss the inferences/limitations of these approaches. The presentation will furthermore survey the currently implemented methods and make recommendations for best practice.

Friday April 17, 2015
2:15 PM - 3:45 PM, Renaissance Ballroom, 5th Floor, Coordinated Session, D6

Two Approaches to Game Based Assessments: Mods and Originals

Session Chair: Andreas Oranje, Educational Testing Service
Session Discussant: Greg Chung, CRESST/UCLA

Demonstration and Brief Introduction
Andreas Oranje, Educational Testing Service

There are two general approaches to game based assessment: modifications and originals. In this symposium we will introduce both, showcasing actual implementations, and provide contrasting views in terms of game and assessment design, analysis approaches, and evaluating efficacy using a range of empirical studies (playtesting through operational data collections).

Modified and Original Games: A Cognitive Science Perspective on Learning and Assessment
Tanner Jackson, Malcolm Bauer, and Masha Bertling; Educational Testing Service

Our current work has compared two approaches for educational game development, original versus modified. We will focus on the differences of these approaches from a cognitive and learning perspective by discussing their impacts on constructs, evidence, scoring, feedback, interpretations of actions, and challenges for learning and formative assessment design.

Modified and Original Game Based Assessment: A Game Design Perspective
Seth Corrigan, Erin Hoffman, and Michael John; GlassLabGames

There is interest in identifying a design approach for game based assessments (GBAs). The proposed presentation details the authors' experiences negotiating constraints posed by two classes of game based assessments when using evidence-centered design. An argument is made for using somewhat different design processes for the two classes of games.

Analysis Approaches for Modified and Original Game Based Assessments
Kristen DiCerbo, Pearson; Katherine Castellano, Helena Jia, Robert Mislevy, and Johnny Lin, Educational Testing Service

This paper explores how differences in game elements lead to differences in the specification of evidence models. Two games will be compared in terms of both the identification of evidence from log files and final results, as well as the accumulation of that evidence using psychometric modeling.

Modified vs. Original: Evaluating Game-Based Assessment Design and Learning
Britte Cheng, Terry Vendlinski, John Murray, and Geneva Haertel, SRI

This paper puts forth a model for the evaluation of game-based assessments that distinguishes between game-based assessments developed "from scratch" vs. those based on "modifications" of an existing game to serve an assessment function.

Friday April 17, 2015
2:15 PM - 3:45 PM, Seville Ballroom East, Lobby Level,
Coordinated Session, D7

Methods for Comparing NAEP Frameworks to Other Assessments and Standards

Session Chair: Markus Broer, Principal Psychometrician/Statistician, American Institutes for Research

Session Discussant: George Bohrnstedt, American Institutes for Research

This set of presentations describes the methodology used for a series of studies comparing the National Assessment of Educational Progress (NAEP) frameworks and item pools with the Trends in International Mathematics and Science Study (TIMSS), the Program for International Student Assessment (PISA), and the Next Generation Science Standards (NGSS).

Historical Overview of Innovations in NAEP Assessment Framework Comparison Studies

Maria Stephens, Kim Gattis, Teresa Neidorf, and Young Yee Kim, American Institutes for Research

This presentation will review the methods used in a series of studies between 2001 and 2013 to compare national and international assessment frameworks and item pools, highlighting the earliest methods used, describing the types of comparisons undertaken (e.g., frameworks, items), and previewing some of the key challenges in this work.

Comparing NAEP and TIMSS Assessment Frameworks in Mathematics and Science

Teresa Neidorf and Kim Gattis, American Institutes for Research

This presentation describes the methods used in a 2012-13 study comparing the mathematics and science frameworks for the 2011 NAEP and Trends in International Mathematics and Science Study (TIMSS) assessments. In addition to quantitative analyses of framework similarity, a qualitative component was introduced to describe similarities and differences in content.

Comparing NAEP and PISA Mathematics Assessment Frameworks and Item Pools

Kim Gattis, Maria Stephens, and Young Yee Kim, American Institutes for Research

This presentation describes the methods used in a 2013-14 study comparing the mathematics assessment frameworks and item pools from NAEP 2013 and PISA 2012. It will focus primarily on a new qualitative component, which provided a systematic side-by-side review of the features of NAEP and PISA items.

Comparing NAEP and the Next Generation Science Standards (NGSS)

Teresa Neidorf, Austin Lasseter, Maria Stephens, and Kim Gattis, American Institutes for Research

This presentation describes the methods used in a 2014-15 study comparing the NGSS and the NAEP assessment frameworks in science, technology and engineering literacy (TEL), and mathematics. It will focus on design considerations and methods used to address the multiple dimensions of the NGSS and the three NAEP frameworks.

Friday April 17, 2015
2:15 PM - 3:45 PM, Seville Ballroom West, Lobby Level,
Coordinated Session, D8

Pseudo Equivalent Groups Linking in Large Scale Assessment

Session Chair: Terran Brown, Educational Testing Service
Session Discussant: Neil Dorans, Educational Testing Service

Shelby Haberman, Hongwen Guo, Junhui Liu, Shameem Gaj, Hyeonjoo Oh, Lu Ru, and Nuo Xi, Educational Testing Service, Princeton, NJ

The goal of the symposium is to introduce the pseudo-equivalent groups linking method by discussing its theoretical background and presenting several applications to large-scale assessment programs. The PEG approach uses adjustment by minimum discriminant information (Haberman, 1984, 2013) and can be applied to the circumstances preventing satisfactory equating requirements.

Theoretical Background of Pseudo-Equivalent Groups (PEG) Linking
Shelby Haberman, Educational Testing Service

This paper introduces a comprehensive theoretical review of PEG linking and procedures to conduct such linking. This paper examines conditions under which pseudo-equivalent groups behave as actually equivalent, and discusses the impact of incomplete background information.

Comparison of PEG Linking With NEAT Equating
Ru Lu and Hongwen Guo, Educational Testing Service

This paper compares the PEG linking results with non-equivalent groups anchor test (NEAT) equating results under different equating situations (i.e., comparison of random equivalent-groups equating and PEG linking; comparison of PEG linking, NEAT linking, and PEG-EAT linking on the sample with large ability difference and less satisfactory anchors).

Application PEG Linking for Testing Mode Adjustment in K-12 Assessment
Hyeonjoo Oh, Junhui Liu, and Shameem Gaj, Educational Testing Service

The purpose of the study is to evaluate the mode comparability of online and paper versions of the K-12 testing program, where no previous scaled scores are available and groups are not equivalent, using the PEG method.

A PEG Linking Study of Matching Variables
Xi Nuo, Hongwen Guo, and Hyeonjoo Oh, Educational Testing Service

This study applies the PEG method to link two tests of different length. Background variables (i.e., gender, ethnicity, grade) as well as the scores on the common items were considered as the matching variables. Linking results using different sets of the matching variables will be discussed in the presentation.

Friday April 17, 2015
2:15 PM - 3:45 PM, Toledo, 5th Floor, Coordinated Session, D9

Reliability, Internal Consistency, and Unidimensionality Related but Distinct Concepts

Session Chair: Ernest Davenport, Jr. University of Minnesota
Session Discussant: Steven Culpepper, University of Illinois at Urbana Champaign

Presenters: Mark Davison, Kyungin Park, and Ernest Davenport, Jr., University of Minnesota

The proposed coordinated session shows reliability, internal consistency, and unidimensionality to be separate but related test attributes. Some novel and other known indices are proposed for the separate concepts. The main contribution of this session is the derivations of functional relationships between these separate indices and hence these separate concepts.

Friday April 17, 2015
2:15 PM - 3:45 PM, Valencia, Lobby Level, Coordinated Session, D10

Beyond Scoring: Alternative Use of Automated Systems for Language Assessments

Session Chair: Su-Youn Yoon, Educational Testing Service
Session Discussant: Christy Schneider, CTB/McGraw Hill

Larry Davis, Su-Youn Yoon, Nitin Madnani, Aoife Cahill, Klaus Zechner, Yu Sun, Educational Testing Service; Vincent Kieftenbeld, McGraw-Hill Education CTB; Lin Gu, Lei Chen, and Zhen Wang, Educational Testing Service

We will discuss the use of automated technology to support a wide range of different processes within a language assessment while improving the reliability and validity. In particular, we will discuss use of automated systems for test development, human score monitoring, and feedback and score report generation.

Using Automated Methods to Identify Overly Similar Discrete Items
Nitin Madnani and Aoife Cahill, Educational Testing Service

Automated machine learning methods are used to identify verbal discrete items that are overly similar to each other and can therefore compromise test validity and reliability if they are included in the same test form.

Monitoring Human Ratings With an Automated Scoring System
Vincent Kieftenbeld, McGraw-Hill Education CTB

This presentation describes different methods that have been developed to monitor human ratings with an automated scoring system. We compare the efficacy of methods based on predicted score and predicted class probabilities with models that were trained specifically to predict from response features whether a response should be reviewed.

Monitoring Human Raters Using Machine Scoring of Spoken Responses
Zhen Wang, Klaus Zechner, and Yu Sun, Educational Testing Service

Automatic scoring systems for constructed response items have the potential to provide solutions to some of the obvious shortcomings in human scoring (e.g., rater inconsistency; rater drift; inefficiency). We recommend using multiple procedures (statistics & plots) to identify "outlier" human raters.

Supplementing Holistic Scores of Speaking With Automated Feedback
Larry Davis, Lin Gu, and Lei Chen, Educational Testing Service

This presentation describes initial efforts to augment holistic scores of English speaking ability with detailed information from an automated scoring engine, to provide feedback to learners taking a practice speaking test. User reactions to a demo feedback instrument will be described, along with considerations for providing automated feedback on unconstrained speech.

Friday April 17, 2015
4:05 PM - 6:05 PM, Renaisaance Ballroom, 5th Floor, Invited Session, E1

Contemporary Problems in Educational Measurement (Satirical Session)

Director: Stephen G. Sireci, University of Massachusetts Amherst
Moderator: Kevin Sweeney, The College Broad

Solving 22nd-Century Measurement Problems
Ellen L. Ripley, Nostromo Inc., Robert Neville, U.S. Department of Education, Elroy Jetson, Spacely Space Sprockets, and Christopher Pike, NASA

An NCME Invited Debate: Godzilla vs. Fairtest: The Rematch
Anne T. Exam, Fairtest; Dr. Godzilla, University of Tokyo

Joint Committee on Fair Testing Practices
David Williamsdaughter, Acid Tests, Inc., Kristen Puff, Regis Philbin Research Fund, Neal Kingdomcum, Yonkers University, Ellen Fortress, misCount, LLC, and Ric Elect, University of North Antarctica

Certifying Psychometric Competence
Andrew Wiley, Alpine Testing and K.T. Han, Council of Cheapskate School Officers

Detecting and Prosecuting Cheaters on Educational Exams
Ellwood U. Cheet, Jake K. Opy, Joliet Correctional Facility, and Robert Crook, Bored of Medical Examiners

Assessing College Readiness: Noncognitive Factors
Gil Andromeda, Even Higher Education Research Consortium, Mary Petunia, Educational Testy Service, Highfive Elephantmat, Professional Procrastination Service, and Sparky Torres, PARCC Inc. Lot

Friday April 17, 2015
6:30 PM-8:00 PM, Seville Ballroom, Lobby Level, InterContinental Hotel

NCME and AERA Division D Joint Reception

Saturday, April 18, 2015
8:00 AM-9:00 AM, Grand Ballroom Salon II, 7th Floor

Chicago Marriott Downtown Hotel (across the street from the InterContinental Hotel)

2015 NCME Breakfast and Business Meeting (ticketed event)

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

Saturday April 18, 2015
9:00 AM-9:40 AM, Chicago Marriott Downtown Hotel,
Grand Ballroom Salon II, 7th Floor, Invited Session

Presidential Address: Educational Measurement: What Lies Ahead

Lauress Wise
HumRRO, Seaside, CA

Improving the effectiveness of our educational systems is vital to providing opportunities for each individual and also to solving the many technical and social problems currently confronting mankind. Measuring student progress is essential to evaluating and improving our educational systems. NCME plays a vital role in building and effectively using assessments of student progress, individually and as a whole. The address will outline some of the key challenges facing our field, including better ways of describing the meaning of reported scores, more precise normative information, measures of more complex skills, and useful and accurate diagnostic information.

Saturday April 18, 2015
10:35 AM - 12:05 PM, Empire Ballroom, 7th Floor, Coordinated Session, F1

Using Ordered Probit Models to Reconstruct Coarsened Test-Score Distributions

Session Chair: Andrew Ho, Harvard Graduate School of Education
Session Discussant: J.R. Lockwood, Educational Testing Service

This symposium investigates the use of heteroskedastic ordered probit models to recover test-score means and standard deviations when only ordinal proficiency data are available. The papers address methodological issues useful to those applying the method and new findings resulting from application of the method to large-scale test-score datasets.

Recovering NAEP and State Test Score Distributions Using Coarsened Data
Benjamin R. Shear, Sean F. Reardon, Stanford University; Katherine E. Castellano, Educational Testing Service; and Andrew D. Ho, Harvard Graduate School of Education

This study evaluates the use of heteroskedastic ordered probit (HOP) models to recover means and standard deviations of real test-score distributions based only on ordinal data from "proficiency"-type categories. We find strong agreement between HOP estimates and estimates based on full continuous distributions, supporting our proposed use.

The Relevance of Normality Assumptions in Ordered Probit Models
Katherine E. Castellano, Educational Testing Service and Andrew D. Ho, Harvard Graduate School of Education

Ordered probit models assume that conditional distributions are respectively normal: normalizable under a common transformation. We evaluate whether real-world test-score distributions meet this assumption. Although we can often reject the null hypothesis that distributions are respectively normal, we find that the violation of assumptions has little impact on parameter recovery.

Practical Strategies for Improving Heteroskedastic Ordered Probit Model Estimates
Andrew D. Ho, Harvard Graduate School of Education; Erin M. Fahle, and Sean F. Reardon, Stanford University

In cases where group sizes are small, fitting heteroskedastic ordered probit models to coarsened test score data can result in biased and imprecise estimates of group standard deviations. This paper introduces and evaluates practical strategies for reducing this bias.

Ordinal Estimation of District Intraclass Correlations in 50 States, 2009-2012
Erin M. Fahle and Sean F. Reardon, Stanford University

Intraclass correlations (ICCs) provide information regarding the amount of variation in test score performance between school districts that is important for study design and policy development. This paper estimates state-level, between-district ICCs of standardized test scores in math and reading in grades 3-8, using ordinal proficiency data.

Saturday April 18, 2015
10:35 AM - 12:05 PM, Exchange, 11th Floor, Coordinated Session, F2

Recent Advances and Comparisons of Teacher Effectiveness Models

Session Chair: Jennifer Broatch, Arizona State University
Session Discussant: Audrey Amrein-Beardsley, Arizona State University

This session will reflect on the impact of various value-added modeling choices from an economic, statistical and educational policy perspective. We will review teacher effectiveness models, specifically value-added models and student growth models, present innovative multidimensional modeling developments and apply the models to evaluate teachers and professional development programs.

Incorporating "Real World Outcomes" in Value-Added Models (VAMs)
Jennifer Broatch, Arizona State University and Jennifer Green, Montana State University

Researchers will present an application of an innovative multidimensional value-added model to assess the impact of a teacher or program on "real-world outcomes" in addition to traditional standardized test scores. This multidimensional model produces reliable and innovative estimates of teacher effectiveness that are in better alignment with educational goals.

Value Added Analysis for Multiple Competencies
Joniada Milla, Sébastien Van Bellegem, CORE; and Ernesto San Martín, Pontificia Universidad Católica de Chile and CORE

In this presentation, researchers use a unique Columbian dataset to estimate multivariate value-added model for several subjects in tertiary education. Researchers then aggregate the estimates to produce a comprehensive index for the overall school value-added in all subjects.

Using Value-Added Models to Assess Teacher Professional Development Programs
Jennifer Green, Montana State University, Jennifer Broatch, Arizona State University, and Walt Stroup, University of Nebraska-Lincoln

Value-added models (VAMs) are often used to provide an estimate of teacher impact on student outcomes. This study explores an extension of VAMs when the goal is to estimate program impact on teacher effectiveness and compares how these estimates differ from standard VAM-based estimates.

Student Growth Percentile: Testing for Validity and Reliability
Margarita Pivovarova and Audrey Amrein-Beardsley, Arizona State University

Researchers analyzed three years of data to evaluate the performance of student growth percentile model for its validity and reliability. They found that past growth measures and observational scores are poor predictors of teachers' current performance. This suggests that growth measures alone do not perfectly capture the underlying teacher quality.

Saturday April 18, 2015
10:35 AM - 12:05 PM, Grand Ballroom, 7th Floor, Coordinated Session, F3

A Potentially Potent Assessment-Literacy Initiative: Reactions Sought

Session Chair: Lou Fabrizio, North Carolina Department of Education

Presenters: James Popham, University of California Los Angeles, Lou Fabrizio, North Carolina Department of Education, Sharyn Rosenberg, National Assessment Governing Board, Holly Spurlock, National Center for Education Statistics, Rebecca Gagnon, National Assessment Governing Board and David Hoff, Hager Sharp

The National Assessment Governing Board, in collaboration with the National Center for Education Statistics, is undertaking a major effort to promote assessment literacy for parents, policymakers, and students. This interactive symposium will describe the project and will present contemplated implementation options. During small-group and large-group discussions, attendees will supply reactions.

Saturday April 18, 2015
10:35 AM - 12:05 PM, King Arthur, 3rd Floor, Invited Session, F4

Session Chair: Li Cai, UCLA
Session Discussant: Howard Wainer, NBME

NCME Career Award Presentation: Item Response Theory, Serendipity, and Bad Questions
David Thissen, University of North Carolina, Chapel Hill

In the first part of this presentation I disclose the role of serendipity in my career, with allusions to the virtues of collaboration, continuing education, and openness. Then I discuss some questions often asked in ways that invite black-and-white, yes-or-no responses, when the right answers are in shades of gray.

Saturday April 18, 2015
10:35 AM - 12:05 PM, Renaissance Ballroom, 5th Floor, Paper Session, F5

Setting Cut Scores

Session Chair: Brian French, Washington State University
Session Discussant: Marianne Perie, Center for Educational Testing and Evaluation

Seven Methods for Estimating Angoff Cut Scores With IRT
Adam Wyse, The American Registry of Radiologic Technologists, St. Paul, MN

This article illustrates seven different methods for estimating Angoff cut scores using IRT models. These include the MLE, EAP, MAP, and WMLE estimators, as well as approaches that have been commonly used based on translating ratings through the test characteristic or item characteristic curves.

Cut Score Estimation: Comparing Bayesian and Frequentist Approaches
Tia Sukin, Pacific Metrics, Dan Segall, DMDC and Alan Nicewander, Pacific Metrics

One criticism espoused for Angoff-based standard setting methods is cut score bias at the extreme cut points. This study explores the statistical differences between cut scores resulting from three cut score estimation methods, consisting both of Frequentist and Bayesian approaches. Preliminary results show similar cut scores result from both approaches.

Establishing Meaningful Expectations for Test Performance via Invariant Latent Standards
Greg Hurtz and Ross Brown, PSI Services LLC

In applied measurement contexts, both test-takers and decision-makers need to comprehend expectations regarding test-taker performance. Setting latent standards that are invariant to specific test content helps to define such expectations across key competence levels. We demonstrate this process, and compare quantitative methods for setting latent standards from standard setting ratings.

Cutscore Distribution Theory (CDT): A Comparison With G-Theory
William Skorupski, Yang Zhao, Joseph Fitzpatrick, and Feng Chen, University of Kansas, Lawrence, KS

A psychometrics for Angoff standard setting is introduced which accounts for panelist consistency and accuracy. The effects of these on resulting cutscores is demonstrated using simulated and real data. Results are compared with a G-Theory approach, which models rater consistency, but doesn't account for panelist accuracy.

Saturday April 18, 2015
10:35 AM - 12:05 PM, Seville Ballroom West, Lobby Level,
Coordinated Session, F6

Psychometric Considerations for the Next Generation of Performance Assessment

Session Chair: Charlene Tucker, K-12 Center at the Educational Testing Service
Session Discussant: Ronald K. Hambleton, University of Massachusetts-Amherst

Presenters: Tim Davey, Educational Testing Service, Steve Ferrara, Pearson, Noreen Webb, University of California-Los Angeles, and Lauress Wise, HumRRO

A distinguished study group of psychometricians worked over the past year to explore psychometric challenges and opportunities presented by the integration of performance assessment into mainstream K-12 assessment systems by the state assessment consortia. Their analysis and recommendations are ready to be shared in the form of four related papers.

Saturday April 18, 2015
10:35 AM - 12:05 PM, Toledo, 5th Floor, Paper Session, F7

Equating Approaches/Methods

Session Chair: Ha Phan, Pearson
Session Discussant: Alina von Davier, Educational Testing Service

Further Study of the Choice of Anchor Tests in Equating
Tammy Trierweiler, Prometric, Charles Lewis, Educational Testing Service, and Robert Smith, Smith Consulting

In this study, we show that the true score correlation between an anchor test and total test is maximized when the anchor test and total test TCCs are proportional, and that, for a fixed anchor TCC, the error variance of the anchor is maximized when the items are equivalent.

MIRT Observed and True Score Equating for Passage-Based Tests
Kyung Yong Kim, Euijin Lim, and Won-Chan Lee, The University of Iowa, Iowa City

The main purpose of this study is to compare the results of IRT observed and true score equating for passage--based tests from five different models using a simulation study. The equating performance of the five models varies as the degree of local dependence changes for items within passages.

The Impact of Anchor Item Embedding Designs on Scale Stability
Shu-Ren Chang, Illinois State Board of Education, Springfield, IL

The purpose of this study is to investigate the impact of difficulty levels and distributions of anchor items on cut-score stability that affects proficiency decisions for examinees/candidates. Eighteen possible anchor embedding designs were investigated. Results provided practical guidelines for practitioners to enhance scale stability and testing fairness.

Comparison of MIRT Equating Procedures for Technology-Enhanced Assessments
Jaime Malatesta, Pearson Education and Won-Chan Lee, University of Iowa

The purpose of this study was to compare the performance of several unidimensional and multidimensional item response theory ((M)IRT) observed-score equating methods using mixed-format tests that contain multiple-choice, free-response, and technology-enhanced item types. The specific equating methods considered were the UIRT, simple structure-MIRT, Bifactor-MIRT, and full-MIRT.

Saturday April 18, 2015
10:35 AM - 12:05 PM, Valencia, Lobby Level, Paper Session, F8

CAT for Diagnostic Purposes

Session Chair: J.P. Kim, ACT
Session Discussant: Kirk Becker, Pearson VUE

Diagnosing Sources of Mathematics Difficulty With Multistage Adaptive Testing
Susan Embretson, Hea Won Jun, and Kristin Morrison, Georgia Institute of Technology, Atlanta, GA

A multistage adaptive testing system for diagnosing sources of mathematical difficulty in middle school is described. The multicomponent latent trait model for diagnosis was applied to year-end tests to diagnose proficiency in four mathematics areas, followed by an adaptive second stage. Results on mastery of areas and skills are presented.

High-Efficiency Item Selection Algorithms for Cognitive Diagnostic Computerized Adaptive Testing
Lei Guo, Faculty of Psychology, Southwest University; Chanjin Zheng, and Hua-hua Chang, Department of Educational Psychology, University of Illinois at Urbana-Champaign

This study proposes two new high-efficiency item selection algorithms, PWC and PWAC, for cognitive diagnostic computerized adaptive testing based on the item discrimination indices developed by Henson and his associates (2005, 2008). They are more efficient than the PWKL index.

Effects of Calibration Error in Cognitive Diagnosis Computerized Adaptive Testing
Hung-Yu Huang, University of Taipei, New Taipei City, Taiwan

In this study, the influence of item calibration error on attribute estimation in cognitive diagnosis computerized adaptive testing was investigated. Under the framework of log-linear cognitive diagnosis model, two restrained models were used to simulate data. The preliminary results showed that the larger error variance the lower correct classification rate.

Effects of Attribute Balancing on Test Efficiency in CD-CAT
Chia-Ling Hsu, Wen-Chung Wang, Assessment Research Centre, the Hong Kong Institute of Education; and Shu-Ying Chen, National Chung Cheng University, Taiwan

Using simulations, this study investigated how attribute balancing affects test efficiency in fixed-length and fixed-precision computerized adaptive testing with cognitive diagnosis models. It was found that attribute balancing had different effects on test efficiency for different item selection methods and termination criteria.

The Variable-Length Adaptive Diagnostic Testing
Yuehmei Chien, Chingwei Shin, Pearson; and Ning Yan, Independent Consultant

This research evaluates different adaptive diagnostic algorithms, focusing on variable-length testing. This has been done by exploring a number of different adaptive item selection algorithm based on different statistical information functions and termination rules. The performance of those adaptive-diagnostic item selection algorithm is evaluated through simulation.

Saturday April 18, 2015
12:25 PM - 1:25 PM, Camelot, 3rd Floor,
Electronic Board Session, Paper Session, G1

Electronic Board #1
Developing and Validating a Self-Report Measure of Attitudes Towards Errors
Jacqueline Leighton, Wei Tang, and Qi Guo, CRAME/University of Alberta, Edmonton, Canada

Students' acceptance and use of formative assessment feedback is impeded by negative attitudes towards their errors. Instruments to measure attitudes towards errors are currently lacking. The present study was designed to develop and begin to validate a self-report measure encompassing the affective, cognitive, and behavioral aspects of attitudes towards errors.

Electronic Board #2
Deterministic, Gated IRT Model for Continuous Probability of Item Cheating
Luyao Peng, University of California Riverside, Riverside, CA

Deterministic, Gated Item Response Theory Model (DGM, Shu, 2013) is used to detect the instances of test cheaters. This study uses DGM to identify test cheaters by incorporating into the model the factor that test items can have continuous probability of being cheated depending on its beta parameter.

Electronic Board #3
The Performance of Five Reliability Estimates at Multidimensional Test Situations
Shuying Sha and Terry Ackerman, University of North Carolina at Greensboro, Greensboro, NC

This study investigates the estimation biases of Cronbach's alpha, and other four reliability indexes by manipulating the following factors: test dimensionality, ability distribution and test discrimination. Results showed that biases increased when true reliability decreased, and those lower bounds tended to overestimate true reliabilities when true reliabilities are low.

Electronic Board #4
A Validity Study Comparing the iPad- vs. Computer-Based Test
Guangming Ling, Educational Testing Service, Princeton, NJ

In this study, we tested 403 8th graders randomly under one of the three conditions: the desktop computer, the iPad alone, and the iPad with an external keyboard. We found that the third condition resulted in longer and better (expected) essays, with no difference on the reading or math scores.

Electronic Board #5
Score Distances of Technology Enhanced Items
Wenhao Wang and Jessica Loughran, University of Kansas, Lawrence, KS

With the advent of the Common Core State Standards, state assessments now include more technology-enhanced (TE) item types. TE items are usually with equal observed score category distance. However, the latent score distances might not be equal. This study aims to identify TE items without equal latent score distances.

Electronic Board #6

Item Position Effects are Moderated by Changes in Test-Taking Effort

Sebastian Weirich, Christiane Penk, Martin Hecht, and Alexander Roppelt, Institute for Educational Quality Improvement (IQB), Humboldt-Universität zu Berlin, Berlin, Germany

The study examines the interdependency of item position effects and test-taking effort. We found that the current test-taking effort diminishes substantially during the test. Position effects are more pronounced for persons with lower initial effort and for persons whose test-taking effort declines in a more pronounced way.

Electronic Board #7

Variability in Proficiency Rates Due to Discreteness in Score Scales

Ying Lu, Educational Testing Service, Princeton, NJ

NCLB has raised immense concerns of the volatility in the percentage of examinees scoring at or above proficient on state assessments. One source of unreliability in percent proficient is discreteness in score scale. This paper suggests an alternative method to calculate percent proficient to capture the longitudinal trend more effectively.

Electronic Board #8

Neuroscience Computing Validates Theory-Based Artistic Judgment Aptitude Construct

Nikolaus Bezruczko, Indiana University Health, Chicago, IL

Cognitive test models emphasize relations between item responses and mental structures yet validation is a major challenge. This presentation expands grounds for mental test validation by corroboration with neuroscience. An artistic judgment aptitude construct was validated with sMRI brain scans, which presented right brain lateralization consistent with theoretical predictions.

Electronic Board #9

The New Psychometric Entrance Test – A Multidimensional Validity Analysis

Dvir Kleper and Noa Saka, National Institute for Testing and Evaluation, Jerusalem, Israel

To explore the reliability, validity and internal structure of the new Psychometric Entrance Test, a Confirmatory Factor Analysis was performed. The results show a good fit of the confirmatory model. A revised model, with additional relationships between scales and factors, shows a better fit compared to the standard model.

Electronic Board #10

An Unfolding-Type Polychoric Correlation

Justin Kern and Chang Hua-Hua, University of Illinois at Urbana-Champaign, Champaign, IL

The polychoric correlation is useful as a coefficient of association between two ordered categorical variables. When the latent response is not assumed to be monotonically related to the observed response, then the polychoric correlation is not appropriate. Here, an alternative assuming a single peak is discussed.

Electronic Board #11

Developing Quality Control Procedures for Continuously Administered Tests

Avi Allalouf and Tony Gutentag, NITE, Jerusalem, Israel

This study deals with the development of new quality control procedures for continuously administered tests, as opposed to traditional test administration modes. It is based on four years of data from examinees who took an online test. It also presents an automated system constructed on the basis of the findings

Electronic Board #12
Receiver Operating Characteristic: A Standard-Setting Tool for Predictive Assessments
Katie Larsen McClarty, Matthew Gaertner, and Daniel Murphy, Pearson

Emphasis on college and career readiness (CCR) assessments has led to new standard-setting approaches that focus on predictive measures. This paper applies Receiver Operating Characteristic (ROC) analysis, commonly used in medical diagnosis, to CCR standard setting. ROC provides multiple indicators, highlighting the trade-offs associated with raising or lowering cut scores.

Electronic Board #13
Exploring Process Data from Problem Solving Items Using Sequence Mining
Qiwei He and Matthias von Davier, Educational Testing Service, Princeton, NJ

This study draws on process data collected in problem-solving tasks in PIAAC to address how sequences of actions are related to task performance. Based on robust n-gram indicators of sequence data identified in different performance groups, a cluster analysis is conducted to examine which patterns of indicators predict group variances.

Electronic Board #14
Using Person-Fit Statistics to Investigate the Effect of Differential Motivation
Marie-Anne Mittelhaeuser, Cito; Wilco Emons, Tilburg University; Anton Beguin, Cito; and Klaas Sijtsma, Tilburg University

If the stakes in testing are low, students may care little whether their scores accurately reflect their maximum performance level. We investigated the difference between responding in low-stakes and high-stakes administration conditions in relation to performance and response consistency. Students differing on account of both consistency and performance were rare.

Electronic Board #15
Detection and Mitigation of Low Motivation on K12 Tests
Elizabeth Stone and J.R. Lockwood, Educational Testing Service, Princeton, NJ

The K12 accountability context has low stakes for individuals and higher stakes at the aggregate (e.g., for teachers or districts). This study evaluated several response-time indices of motivation for a test aligned with the 7th-grade Common Core State Standards for mathematics for students with and without learning disabilities.

Electronic Board #16
Not Just Free Lunch: A Neighborhood-Based SES Variable for Districts
Christopher Moore, Luke Stanke, Eric Vanden Berk, Amanuel Medhanie, Minneapolis Public Schools; and Martin VanBoekel, University of Minnesota

Many stakeholders use free lunch eligibility as a proxy for socioeconomic status (SES). The goal of this study is to develop a better measure of SES for school districts. Using geographic methods, this study combines student-level information with data from the American Community Survey to create an SES variable.

Electronic Board #17
Device Comparability of Tablets and Computers for Assessment Purposes
Laurie Davis and Yuanyuan McBride, Pearson, Austin, TX

The definition of "computer-based testing" is becoming more nuanced as BYOD and 1:1 programs increase the use of tablets and other devices in classrooms. In this paper researchers explore device comparability through a large scale quantitative study of the differences in student performance when testing on tablets and computers.

Electronic Board #18
Evaluating Properties of Scores on Mixed-Format Tests Using IRT
Won-Chan Lee, Jiwon Choi, Yujin Kang, and Stella Kim, University of Iowa, Iowa City, IA

Test scores on mixed-format tests are evaluated in terms of various psychometric properties including conditional standard errors of measurement, classification consistency/accuracy, and reliability. Three different IRT frameworks--unidimensional, bifactor, and simple structure-will be considered and compared using several real mixed-format tests with different levels of item-format effects.

Electronic Board #19
Modeling English Proficiency Growth When Reclassification is Informative
Tyler Matta, Meg Guerreiro, University of Oregon; and Moti Hara, Portland State University

A challenge in estimating language attainment trends for Limited English Proficient students is that reclassification to full English proficiency results in non-ignorable attrition. This paper presents an approach for producing unbiased growth estimates by modeling the joint distribution of longitudinal English proficiency scores and time-to-reclassification simultaneously using a shared-parameter model.

Electronic Board #20
Does the NAEP Model Adequately Predict the Achievement Gap?
Matthew Johnson, Teachers College, Columbia University and Sandip Sinharay, Pacific Metrics

In this talk we use Bayesian posterior predictive checks to examine the appropriateness of the normal linear regression model assumed by NAEP. In particular we examine whether the NAEP model adequately explains the summary statistics of important demographic groups and relevant measures of the achievement gaps between the groups.

Electronic Board #21
Gathering Evidence of Response Processes for Alternate Assessments (AA-AAS)
Russell Swinburne Romine, Amy K. Clark, and Meagan Karvonen, University of Kansas, Lawrence, KS

Validity arguments commonly use cognitive labs as one source of evidence about student response processes. However, there are challenges in collecting such evidence for alternate assessments designed for students with significant cognitive disabilities (AA-AAS). We present findings from cognitive labs and test administration observation sessions for an AA-AAS.

Electronic Board #22
The Reduced RUM as a Logit Model: A Demonstration via Mplus
Chia-Yi Chiu, Rutgers, The State University of New Jersey and Hans-Friedrich Koehn, University of Illinois at Urbana-Champaign

Commercial implementations of the EM algorithm for fitting the Reduced RUM are available in the LCA routines of Latent GOLD or Mplus, for example. In this proposal, the general parameterization of the Reduced RUM as a logit model and the associated parameter constraints are derived.

Electronic Board #23
Empirical Estimates of Student and School Level Variance Components
Jehanzeb Cheema, University of Illinois at Urbana-Champaign, Champaign, IL

Cross-country estimates of student- and school-level variance components of student literacy data are provided. These can be used to (1) correct effect sizes reported in prior studies that ignored nested structure, (2) rank/group countries with respect to variance estimates, which may provide valuable insight into factors responsible for such variation.

Electronic Board #24

Are Multimedia Items More Memorable and Prone to Compromising?

Feiming Li, University of North Texas Health Science Center; Hao Song, and Yi Wang, NBOME

This study demonstrated how to apply the moving average technique to both response and response time for monitoring the potential item compromise on paired multimedia and text items in a computer-based exam. The result will shed light on whether multimedia items are more prone to compromising.

Electronic Board #25

Evaluating the Misalignment Between Information and Content Specifications

M. Fernanda Gándara and Lisa Keller, University of Massachusetts Amherst, Amherst, MA

Ignoring item information functions in test specifications produces a misalignment between the proportions of items and information they provide. This has implications for content representation, as examinees may take tests that are different from an information standpoint. This work discusses the problem and provides a method to evaluate its significance.

Saturday April 18, 2015
12:25 PM - 1:55 PM, Empire Ballroom, 7th Floor, Paper Session, G2

Assessing Diverse Learners

Session Chair: Robert Schwartz, Pearson
Session Discussant: Ellen Forte, edCount

The Effects of Initial ELL Classification on Later Academic Achievement
Nami Shin, UCLA, Los Angeles

This study will explore the effects of initial English Language Learner (ELL) classification on students' later educational experience and academic achievement. The particular focus is on students near the cut-off for Initially Fluent English Proficient (IFEP)/ ELL, that is, students who are just above or just below the cut-off scores.

Assessing the Effect of Language Demand in Math Word Problems
Kathleen Banks, Middle Tennessee State University, Cindy Walker, University of Wisconsin-Milwaukee, and Ahmad Jeddeeni, Middle Tennessee State University

Differential bundle functioning (DBF) analyses were conducted to determine whether seventh and eighth grade second language learners (SLLs) had lower probabilities of answering bundles of math word problems correctly that had heavy language demands, when compared to non-SLLs of equal math proficiency.

Digging Deeper: A Latent Class Analysis of English Learners
Molly Faulkner-Bond, University of Massachusetts Amherst/Educational Testing Service

Latent class analysis is used to create subgroups within a sample of third grade English learners. Classes are analyzed and interpreted with respect to variables such as (current) level of English proficiency, (subsequent) year of reclassification, home language, and immigrant status. Implications for research, policy, and practice are discussed.

Item Construct Maintenance When Varying Levels of Support and Complexity
Anne Davidson, Smarter Balanced Assessment Consortium; Sarah Hagge, Minnesota Department of Health; Bill Herrera, Charlene Turner, edCount; and Martha Thurlow, University of Minnesota

Universal design emphasizes access to test constructs for diverse students. The study investigates an evidence centered-design outcome (item tiers) focused on maintaining test construct while varying accessibility features. Convergent/discriminant analysis evaluates construct maintenance; multiple-sample SEM investigates invariance of test structure across subgroups. Preliminary results suggest that design goals were met.

Saturday April 18, 2015
12:25 PM - 1:55 PM, Exchange, 11th Floor, Paper Session, G3

Automated Scoring and Text Generation

Session Chair: Pu-Wai Lei, Penn State University
Session Discussant: Andre Rupp, Educational Testing Service

Hierarchical Latent Variable Models for Human and E-Rater Scored Responses
Peter van Rijn, Educational Testing Service Global and Mo Zhang, Educational Testing Service

We investigate the latent structure of English language arts assessments that consist of selected response, short constructed response, and essay items. We develop an integrated hierarchical latent variable model to differentiate the impact of item type (SR and CR) and scoring type (human and e-rater) on measurement precision.

Statistical Models for Automated Essay Scoring Engine Training
Scott Wood and Sue Lottridge, Pacific Metrics Corporation, Lakewood, CO

Many statistical models are available for mapping essay features to human-assigned scores in automated essay scoring engines. The purpose of this research project is to understand how metrics of automated essay scoring quality change under different statistical and machine learning models. Metrics include exact agreement rates and quadratic weighted kappa.

Generating Models of Student Writing Abilities Through Text Analysis
William Bryant, ACT and R. Gordon Rinderknecht, University of Maryland

This study describes a method for generating models of writing abilities from student responses to direct writing assessments. Automated text analysis data is correlated with qualitative evaluations. The resulting models provide insight into the characteristics of writing and the progression of writing abilities within and across grades.

Automated Capturing of Psycho-Linguistic Features in Reading Assessment Text
Makoto Sano, Prometric Inc., Baltimore, MD

This study used PLIMAC, a natural language processing tool, to automatically capture psycho-linguistic features of passage based multiple choice items. Items from the 2011 NAEP Grade 8 Reading assessment were evaluated and multiple linear regression analyses were conducted to identify psycho-linguistic features that best predicted overall item difficulty.

Saturday April 18, 2015
12:25 PM - 1:55 PM, Grand Ballroom, 7th Floor, Coordinated Session, G4

Ensuring Content Validity and Alignment of Computer Adaptive Reading Assessments

Session Chair: Craig Mills, McGraw-Hill Education CTB
Discussant: Tim Davey, Educational Testing Service

Presenters: Liru Zhang, Delaware Department of Education; Seung W. Choi, Wim van der Linden, McGraw-Hill Education CTB; Shudong Wang, NWEA; Hong Jiao, and Rosalyn Bryant, University of Maryland

CAT in passage-based reading must address constraints imposed at the item- and passage-levels, as well as complications in reporting categories due to dependency between a passage and associated items. In this session, innovative approaches are investigated; comparability of individual adaptive forms is examined; and alternate item selection methods are explored.

Saturday April 18, 2015
12:25 PM - 1:55 PM, King Arthur, 3rd Floor, Paper Session, G5

Technical Investigation of SGPs/VAMs for Teacher Evaluation

Session Chair: Andrew Mroch, ACT
Session Discussant: Andrew Ho, Harvard Graduate School of Education

Estimating Individual Error Variances for Student Growth Percentiles Under IRT
Jinah Choi and Robert Ankenmann, University of Iowa, Iowa City

This paper investigates using item response theory (IRT) to estimate individual standard errors of measurement (SEMs) for student growth percentiles (SGPs). The simulation study shows a series of processes for generating longitudinal data, estimating individual SEMs, and constructing confidence intervals for SGPs. Reporting of results is also discussed.

Locating Student Growth Projections in a Familiar Regression Framework
Katherine Furgol Castellano, Educational Testing Service, San Francisco, CA

The Student Growth Percentile (SGP) model allows for predicted future scores through student growth projections/trajectories that reflect an array of possible growth scenarios given student past performance. This paper grounds these projections in a familiar parametric regression framework and explicates the effects of their assumptions on predictive accuracy.

Exploring the Impact of Cohort Variability on Teacher Effects
Daniel Anderson and Joseph Stevens, University of Oregon, Eugene, OR

Year-to-year variability in achievement across five student cohorts was explored. Preliminary results suggest students' within-year growth differs significantly by cohort, despite non-significant differences in initial achievement, independent of the teacher to whom students were assigned. Models for teacher effects that do not account for cohort variability may therefore be biased.

Assessment Properties and Value-Added Measurement of Educator Effectiveness
Yang Wang, Education Analytics; Nandita Gawade, and Robert Meyer, University of Wisconsin - Madison; Education Analytics

This paper explores the association between assessment properties and value-added estimates on educator effectiveness. Based on empirical findings from over 100 tests, recommendations are provided to practitioners on developing quality assessments to enable valid and reliable measurement of educator effectiveness in both traditionally tested and non-tested grades and subjects.

Saturday April 18, 2015
12:25 PM - 1:55 PM, Renaissance Ballroom, 5th Floor, Paper Session, G6

Performance Level Descriptors

Session Chair: Jamin Huggins, ACT
Session Discussant: Kuzey Bilir, Pearson

Anchored Graphical Representations: An Alternative to Traditional Performance Level Descriptors
Richard Tannenbaum, Irvin Katz, and Priya Kannan, Educational Testing Service, Princeton

We applied concepts from graphic organizers and frame-of-reference training to develop Anchored Graphical Representations (AGRs) for a teacher licensure test. AGRs situate traditional performance level descriptors (PLDs) within a larger range of performance expectations by content area. AGRs resulted in lower passing scores and less likelihood for panelists to over-generalize.

Incorporating Cognitive Diagnostic Information into the Standard Setting Process
Joe Grochowalski, Fordham University and Leslie Keng, Pearson

We use a cognitive diagnostic model to add pages to an ordered item booklet to mark where performance level descriptors and levels (e.g. basic, advanced) fall. We use actual data and compare the locations of the empirical and expert-placed bookmarks. We discuss optimization of this method to facilitate standard setting.

Developing a Framework for the International Benchmarking of Performance Standards
Andrew Wiley and Susan Davis-Becker, Alpine Testing Solutions, Inc., Brooklyn, NY

The use of international benchmarks during standard setting is becoming increasingly common. While the inclusion of such benchmarks can provide valuable context, the appropriate procedures for using them have not been defined. This paper will present a framework for the use of international benchmarks in standard setting for educational assessments,

Validating Performance Level Descriptors Through a Longitudinal Examination of Data
Mary Hansen, Robert Morris University, Peter Heh, California University of Pennsylvania, Steve R. Lyon, University of Pittsburgh

This paper investigates the match between student response data and Performance Level Descriptors (PLDs) for one state's Alternate Assessment based on Alternate Achievement Standards. Through longitudinal analysis across seven administrations of the science assessment, comparisons of PLDs to actual student performance data are provided using numerical and graphical displays.

Saturday April 18, 2015
12:25 PM - 1:55 PM, Seville Ballroom East, Lobby Level, Paper Session, G7

Irregularities in Operational Testing

Session Chair: Sarah Schnabel, American Board of Ophthalmology
Session Discussant: Leslie Keng, Pearson

Examining Test Irregularities in Mixed-Format Testing
Xin Li, Tianli Li, and Chi-Yu Huang, ACT, Inc., Iowa City, IA

This study entailed a simulation study to assess the performance of different statistical methods in investigating individual- and classroom-level test irregularities when mixed-format tests are administered with the consideration of the possible presence of multidimensionality due to the mixture of item formats.

Getting Lucky: Guessing's Threat to the Validity of Performance Classifications
Brett Foley, Alpine Testing Solutions, Denton, NE

When using multiple choice items, situations exist where guessing at random can be an effective strategy for passing, thus lowering the validity of content-based interpretations of the test results. This study addresses test development decisions that can help avoid/remediate situations where random guessing can be an effective strategy.

Preknowledge Detection Using a Scale-Purified Deterministic Gated IRT Model
Carol Eckerly, University of Wisconsin-Madison, Ben Babcock, American Registry of Radiologic Technologies, and James Wollack, University of Wisconsin-Madison

This paper introduces an iterative, scale purified approach using the Deterministic Gated Item Response Theory Model to identify examinees who have likely benefitted from item preknowledge. This method reduces bias in parameter estimates used in the model for classification of examinees as either those with preknowledge or those without preknowledge.

Identification of Non-Random Missing Responses in Computer-Based Assessments
Andreas Frey and Christian Spoden, Jena University, Jena, Germany

A recursive method for separating item responses which are missing at random from those not missing at random using response times is introduced and is illustrated with multidimensional achievement test data. The method proved to be effective and led to small effects on item and ability parameters.

Saturday April 18, 2015
12:25 PM - 1:55 PM, Seville Ballroom West, Lobby Level, Paper Session, G8

Automated Scoring

Session Chair: Jen Beimers, Pearson
Session Discussant: Peter Foltz, Pearson

Comparing the Performances of Different Machine Learning Methods in Automated Essay Scoring
Jing Chen, James Fife, and Mo Zhang, Educational Testing Service, Princeton

Most automated essay scoring programs use a linear regression model to predict an essay score from a set of computer generated feature scores. This study compares the performances of the linear regression model and some alternative models based on machine learning algorithms in predicting human ratings.

Detection of Aberrant Responses in Automated Scoring
Mo Zhang and Jing Chen, Educational Testing Service, Princeton, NJ

The current state-of-the-art automated scoring field calls for development of defense system for aberrant responses. This study investigates the approaches to the detection of aberrant responses in automated scoring context, and to effective classification of responses into automatedly scorable and non-scorable categories during pre-screening as well as post-hoc screening stages.

IRT-Based Reliability Estimates for Human and Machine Scored Essays
Alan Nicewander, Tia Sukin, and Sue Lottridge, Pacific Metrics, Monterey, CA

This inquiry continues work done by the present authors. In that research, theory was proposed for estimating the reliability of essay scores based on fitting IRT models to these scores. The present study further develops the IRT theoretical basis for estimating the reliability of essay scores and presents real-data examples.

Saturday April 18, 2015
12:25 PM - 1:55 PM, Toledo, 5th Floor, Paper Session, G9

Equating Methods

Session Chair: Jaime Malatesta, Pearson
Discussant: Robert Brennan, University of Iowa

The Use of Poisson-Binomial Distribution in Equating Test Scores
Jorge González, Pontificia Universidad Católica de Chile and Marie Wiberg, Umeå University

The Poisson-Binomial model is proposed for modelling the conditional distribution of scores for a given ability when using equating methods. Using simulations, the proposed model is compared with the well-known Lord & Wingersky (1984) algorithm for compound binomials, showing promising results. Implications for its use in equating are discussed.

An Investigation of Various Approaches to Developing a Theta-Based Scale
Tianli Li, Xin Li, and Troy Chen, ACT Inc., Iowa City, IA

This study investigated various approaches to creating a theta-based scale matching statistical properties of an existing number-correct based scale. While holding a constant CSEM, three direct theta-to-scale approaches using a linear or cubic transformation or an equipercentile method were explored. One indirect approach involving true scores was also examined.

Evaluating the Robustness of Four Equating Methods Under Common-Item Design
Chunyan Liu, Chi-Yu Huang, and NooRee Huh, ACT, Inc., Iowa City, IA

In this simulation study, four equating methods (Tucker, Levine, frequency estimation, and chained equipercentile) under the common item non-equivalent groups design will be compared when various conditions of group/form difference are considered. The results of the study will provide practitioners some guidelines about which method is preferred under what conditions.

Comparison of IRT Based and CTT Based Pre-Equating Approaches
Meichu Fan and Qing Yi, ACT, Inc., Iowa City, IA

Pre-equating research has tremendous appeal to test practitioners with the increasing demand for immediate score reporting. IRT pre-equating research is readily applicable, but research on pre-equating using classical test theory (CTT), where only classical item statistics are available, is limited. This study compares IRT to CTT pre-equating approaches.

Saturday April 18, 2015
12:25 PM - 1:55 PM, Valencia, Lobby Level, Paper Session, G10

Methods for Investigating Threats to Validity

Session Chair: Jie Lin, Pearson
Session Discussant: Paul Nichols, ACT

Developing a Large-Scale Assessment Using Evidence-Centered Design: Did It Work?
Claudia Flowers, UNC Charlotte; Martha Thurlow, Rachel Quenemoen, University of Minnesota; Liz Towles-Reeves, Bill Herrera, Charlene Turner, edCount; Anne Davidson, Smarter Balanced; and Sarah Hagge, Minnesota Department of Health

Evidence-Centered Design (ECD) provides a systematic framework for designing assessments in terms of evidentiary arguments. This presentation reports the results of the ECD assessment implementation phase of a large-scale assessment. Empirical evidence will be used to test conceptual assessment framework. Implications for using ECD will be provided to attendees.

Addressing the Dilemma of Nested Structures in Construct Validity Research
Maria Elena Oliveri and Daniel McCaffrey, Educational Testing Service, Princeton, NJ

We present a framework to guide decision-making regarding the use of a multilevel approach to validity research. Validation arguments must include an explicit articulation of the unit of inference. Hypotheses about constructs must identify the level to which they apply. We demonstrate the framework on the assessment of noncognitve constructs.

Detecting Inattentive Pilot Test Examinees in an MTURK Sample
Avi Fleischer, Illinois Institute of Technology and Alan Mead, Talent Algorithms Inc.

Inattentive pilot test examinees cause problems during item and reliability analyses. This study examines the effectiveness of a validity index strategy, normally employed with personality measures, for detecting and removing inattentive examinees on Math and English pilot exams administered to an MTurk sample.

Holistic Scoring and IRT-Based Classification for Evaluating Learning Progressions
Edith Graf, Educational Testing Service and Peter van Rijn, Educational Testing Service Global

Learning progressions are provisional structures that require empirical verification. A question in examining learning progressions is whether the specification and ordering of levels may be empirically recovered. For this purpose, we compare a holistic scoring approach to a classification approach based on a constrained version of the partial credit model.

Using Bayesian Network Analysis to Validate a Mathematics Learning Map
Anu Sharma, Angela Broaddus, Ayse Esen, and Feng Chen, University of Kansas, Lawrence, KS

The present study examines the validity of a learning map that serves as a framework for developing instructionally embedded assessment tasks. Bayesian analysis is used to assess the learning trajectory in relation to the understanding of patterns. Results were used to refine the mathematical connections depicted in the map.

Saturday April 18, 2015
2:15 PM - 3:45 PM, Camelot, 3rd Floor, Electronic Board
Session: GSIC Student Issues Poster Session, H1

GSIC Electronic Board Poster Session

Graduate Student Issues Committee
Lisa Beymer, Chair
Laine Bradshaw, Jeremy Brown, Laurie Davis, Jerusha Gerstner, Jason Herron, Evelyn Johnson, David King, Ray Reichenberg, and Ting Wang

Electronic Board #1
Model Selection Methods for Passage Based Tests
Euijin Lim, Kyung Yong Kim and Won-Chan Lee, The University of Iowa, Iowa City

The purpose of this study is to compare several model selection methods for choosing appropriate IRT models for passage-based tests. Three IRT models are considered: unidimensional, testlet response, and bifactor models. The performance of AIC, BIC, CVLL, and DIC is compared using a simulation study.

Electronic Board #2
A Method to Allow Item Review in MCAT
Zhe Lin, Ping Chen, and Tao Xin, Beijing Normal University, Beijing, China

Most CATs and MCATs don't allow item review due to deterioration of measurement precision and extra cheating strategies. To avoid these problems, this study proposed successive block with item pocket method to avoid much loss of precision while allowing item review in MCAT. It has well compatibility to all MCATs.

Electronic Board #3
Comparing MULTILOG and IRTPRO Parameter Estimates Under Skewed Trait Distributions
Brian Leventhal, Clement Stone, Lan Yu, and Carol Greco, University of Pittsburgh, UPMC

From an analysis of responses to a PROMIS health assessment, Multilog and IRTPro parameter estimates were found to differ markedly. The purpose of this study was to further explore these differences and better understand the behavior of MML estimation procedures when the trait distribution is highly skewed.

Electronic Board #4
The Effects of Early Acceleration on Students' Academic Achievement
Julia Kretschmann, Miriam Vock, University of Potsdam, Germany; and Oliver Lüdtke, Leibniz Institute for Science and Mathematics Education, Germany

Based on longitudinal data covering 3 measurement occasions, we examined the effects of grade skipping on academic performance. Various types of propensity score matching were attempted and, considering balancing multiple covariates, full matching was applied. Same-grade comparisons indicate that skipped students keep up with their older equally gifted classmates.

Electronic Board #5
Improving Equating Precisions by Incorporating Prior Information of Common Items
Wenchao Ma and Jimmy de la Torre, Rutgers, The State University of New Jersey, New Brunswick, NJ

In common item equating design, the parameter estimates of common items based on the old form can be specified as priors when calibrating the new form. Its performance will be evaluated and a comparison with other equating approaches will be examined using simulations based on real data.

Electronic Board #6
Investigation of Dependence of Equating Methods on Group Difference
Shichao Wang, The University of Iowa, Wei Wang, ETS, and Michael Kolen, The University of Iowa

The purpose of this study is to examine the dependence of various equating methods on group difference under the common-item nonequivalent groups design for mixed-format tests. Pseudo groups with certain performance levels will be used. Both traditional and item response theory equating methods will be examined.

Electronic Board #7
Evaluating Schools With Respect to Growth of Students in Subpopulations
Yixing Liu, Roy Levy, and Nedim Yel, Arizona State University, Tempe, AZ

Bayesian methods and the residual model were proposed to characterize schools with respect to the students' growth in subpopulations (i.e. students with disabilities) within the framework of a three-level model. Five specific methods were compared and the result indicated that correlations among the five methods ranged from .85 to .99.

Electronic Board #8
Classification in MIRT With Subscore Reporting and Reliability Comparison
Keyin Wang and Liyang Mao, Michigan State University, Lansing, MI

This study aims to evaluate the classification accuracy yield from the multidimensional and unidimensional IRT models for a three-dimensional test. In addition, two subscore reliability measures are compared. The effect of the subtest length and correlation among dimensions on classification accuracy and subscore reliability is also examined.

Electronic Board #9
Psychometric Analysis of a Student Risk Assessment for Washington Courts
Jessica Beaver, Brian French, Chad Gotch, Paul Strand, Washington State University; and Carl McCurley, Washington State Center for Court Research

This study refines and increases the accuracy of the Washington Assessment of the Risks and Needs of Students (WARNS) scale. Additional validity evidence provides support for the use of WARNS scores for decisions about individuals. Item analyses, factor analysis, and known-group mean comparisons are presented.

Electronic Board #10
Accounting for Uncertainty of Item Parameter Estimates on Ability Estimates
Ragip Terzi and Jimmy de la Torre, Rutgers, The State University of New Jersey

A new method was proposed to account for some of the uncertainty of item parameter estimates on ability estimates. The statistical inference of ability estimates was based on the confidence interval, incorporating the new standard error of the item estimates. It resulted in accounting for more uncertainty on ability estimates.

Electronic Board #11
A Comparison of Two Models for Hierarchical Item Response Data
Xueying Hu Francis, Texas A&M University-College Station, College Station, TX

Evaluates the estimation accuracy of person and item parameters of Kamata's MLIRT and Multiple Regression IRT models for hierarchically structured data. Investigates the performance of school-level ability variance estimates. Discusses the advantages and disadvantages of each model in accommodating hierarchical item response data. Indicates the practical applications of models.

Electronic Board #12
Omitting Time-Varying Confounders When Predicting Growth Trajectories: A Case Study
Marcus Waldman, Harvard Graduate School of Education, Malden, MA

Residual diagnostics suggest that the omission of time-varying confounders explains up to 90% of predictive misfit in a simple OLS model. The effect of confounding is not only poor predictions, but also unintended policy incentives. We discuss alternative models to mitigate the ill-effects of omitted variable bias.

Electronic Board #13
Differential Item Functioning as an Artifact of Model Misspecification
Nathan D. Minchen, Lokman Akbay, and Jimmy de la Torre, Rutgers, The State University of New Jersey, New Brunswick, NJ

Due to its simplicity, the Rasch model can be a first choice among educational practitioners and researchers. However, this simplicity can also come at a cost. This paper explores the set of conditions under which differential item functioning (DIF) is induced when 3PL data is fitted with the Rasch.

Electronic Board #14
Ability Estimation and DIF Detection in Large-Scale Assessments
Luciana Cancado, Logan Rome and Bo Zhang, University of Wisconsin-Milwaukee, Milwaukee, WI

Large-scale assessments often use a rotated booklet design and many secondary analyses of these assessments require the estimation of ability. This simulation study examines the impact of different ability estimates (total scores, point estimates, plausible values) on DIF detection using logistic regression under varying item overlap (or booklet rotation) conditions.

Electronic Board #15
Modified Maximum Priority Index for Exposure Control in Multidimensional CAT
Liyang Mao, Xin Luo, Michigan State University; and Xuechun Zhou, Pearson

This study identifies the misuse of the Maximum Priority Index (MPI) method for item exposure control in multidimensional CAT and proposes the Modified MPI method. The Modified MPI method can effectively control the item exposure rate without inflating the item pool usage.

Electronic Board #16
The Effects of Dimensionality on Differential Item Functioning Analysis
Yuan-Ling Liaw, University of Washington, Seattle

The effects of "dimensionality" on differential item functioning (DIF) analysis will be explored, particularly with respect to the magnitude of the correlation between the target dimension and nuisance dimension as well as the impact of the primacy of the target dimension.

Electronic Board #17
The Influence of Survey Format on Results
Andrew Ainsworth, David Martinez Alpizar and Scott Plunkett, CSUN, Mission Hills, CA

In an experimental design (n=1527 college students), multigroup SEM demonstrated that responses about mothers and fathers on same page are significantly more correlated than responses about mothers asked on separate page from responses about fathers; thus separate pages are recommended. The results have application to studies on teachers, classrooms, etc.

Electronic Board #18
Comparing Overall Ability Estimation Between Multidimensional and Unidimensional IRT Models
Mingcai Zhang and Lihong Yang, Michigan State University, East Lansing, MI

This paper compared the overall latent trait estimated from different MIRT models and unidimensional IRT model. Different correlation matrices between two latent traits and different sample sizes were examined to check the model performance.

Electronic Board #19
Detecting Aberrant Behaviors With a Hierarchical Lognormal Response Time Model
Zhen Li, University of California, Los Angeles and Jessalyn Smith, CTB McGraw-Hill Education

The goal of this study is to examine the detection rates of aberrant responses under a variety of settings with a hierarchical lognormal response time model (van der Linden, 2006). Results show that empirical data fits the model quite well. Negative and positive aberrant RTs are identified.

Electronic Board #20
Wholistic Scoring of Additive Structures in Constructed Responses
Catherine Kaduk, University of Illinois at Chicago, Naperville, IL

Representations used in students' model-making on a constructed response word problem were analyzed for evidence of students' varying levels of understanding. Problem representation data from third grade classrooms were feature-scored for evidence of increasing levels of structure and understanding. Wholistic and feature scoring approaches are compared.

Electronic Board #21
Links Between Teacher Judgment Accuracy and Differentiated Instruction
Andrea Westphal, Anna Gronostaj, and Miriam Vock, Universität Potsdam, Potsdam, Germany

Hierarchical models were used to test the association between two indicators of teacher judgment accuracy (rank component and level component) and student ratings of differentiated instruction. Results indicate that good judges of students' ability level use more differentiated instruction. Rank judgment accuracy was not associated with differentiated instruction.

Electronic Board #22
Modifying Mantel-Haenzel DIF Detection Hypotheses to Include Effect Size
Phillip Sherlock and Brian Habing, University of South Carolina, Columbia, SC

Uttaro and Millsap's (1994) factorial design was replicated to investigate the ability of an alternative Mantel-Haenzel procedure for DIF detection to control for Type-I error inflation. The modified procedure incorporated effect size measures into the hypotheses to control for the effects of test length and ability differences on DIF detection.

Electronic Board #23

Effect of Q-Matrix Design Under Hierarchical Attribute Structures

Lokman Akbay and Jimmy de la Torre, Rutgers, The State University of New Jersey, New Brunswick

Not all possible attribute patterns are permissible when attributes follow a hierarchy. Consequently, there is a reduced number of possible attribute combinations used in the Q-matrix of Attribute Hierarchy Method. This study investigates the impact of structured and unstructured Q-matrices on CDM parameter estimation and attribute classification under hierarchical cases.

Electronic Board #24

A Simulation Study of Missing Data on IRT Parameter Estimates

AhYoung Shin and Won-Chan Lee, University of Iowa, Iowa City

A simulation study is conducted to compare the effect of conventional and modern ways for handling missing data for educational tests composed of dichotomous items. The study examines how varying proportion of missing data affects the estimation of IRT parameters under different testing conditions.

Electronic Board #25

Comparative Analyses of Popular MIRT Models and Software

Guler Yavuz and Ronald Hambleton, University of Massachusetts Amherst, Amherst, MA

The purpose was to investigate model parameter recovery of two popular MIRT parameter estimation software, BMIRT and flexMIRT; and two multidimensional models, compensatory three parameter logistic model and graded response model; and three item parameter estimation techniques used with these software packages, under some common situations. Practical implications are offered.

Saturday April 18, 2015
2:15 PM - 3:45 PM, Empire Ballroom, 7th Floor, Invited Session, H2

Measurement and Implementation Challenges in Early Childhood Assessment

Session Chair: Michael Rodriguez, University of Minnesota
Session Discussant: Kristen Huff, Regents Research Fund

Growth and Development in Preschool as Foundations for a Modern Measurement Model
Scott McConnell, University of Minnesota

Innovation and modern approaches to measurement in early childhood education must align with existing program functions for wide-scale adoption. This presentation will highlight core assumptions of early education with relevance for measurement development and application. Examples from existing measures, and challenges faced in their development and refinement, will be presented.

Standard Setting for Spanish Individual Growth & Development Indicators
Alisha Wackerle-Hollman and Michael C. Rodriguez, University of Minnesota

This presentation will introduce the Spanish Individual Growth and Development Indicators (S-IGDIs) and provide a detailed summary of the standard setting procedures used to identify benchmarks for performance at fall, winter and spring screening periods. Challenges in the process and innovations in the approach will be discussed.

Multi-Phase Identification of Preschool Students With Behavioral Difficulties: Evaluation Within a Multi-Trait, Multi-Method Framework
Ryan Kettler, Rutgers, The State University of New Jersey

The Preschool Behavior Screening System (PBSS) incorporates multiple phases and raters as a universal screener for students with both internalizing and externalizing problems. A Multi-Trait, Multi-Method Framework was used to evaluate the relative contribution of trait versus informant to the pattern of relationships among scores from the PBSS and BASC-2.

Implementing a Comprehensive Assessment System in Early Childhood: Challenges and Opportunities for Further Growth
Megan Cox, Minnesota Department of Education

Minnesota received Race to the Top-Early Learning Challenge funding to build and enhance early learning programs for children with high needs. This presentation will address requirements of an early childhood comprehensive assessment system, including implementation and strategies to promote early childhood screening, formative assessment, and kindergarten entry assessment initiatives.

Saturday April 18, 2015
2:15 PM - 3:45 PM, Exchange, 11th Floor, Coordinated Session, H3

Issues in Human Scoring of Constructed-Response Items

Session Chair: Edward Wolfe, Pearson
Session Discussant: Isaac Bejar, Educational Testing Service

Presenters: Edward Wolfe, Pearson, Jo-Anne Baird, Oxford University, Lawrence DeCarlo, Teachers College, Columbia University, Michelle Meadows, Ofqual, and Yoav Cohen, National Institute for Testing & Evaluation

This session focuses on rating errors and efforts to minimize them. The papers focus on the literature on factors that produce rater effects; whether rating errors in operational programs should cause concern; how table effects can be minimized via online, distributed training; and the ill-advised nature of adjudication.

A Causal Model of Human Scoring Behavior in Educational Assessments

This literature review summarizes four categories of substantive research regarding the rating process (rater characteristics, response content, rating process, and assessment design) and identifies potential causal links between these features and the emergence of rater effects in assigned scores.

Scoring as Signal Detection: Implications for Rater Effects and Classification

Scoring of constructed response items is placed within a signal detection theory framework. Rater effects (e.g., 'severity', 'central tendency') are shown to arise from the raters' use of response criteria, whereas 'halo effect' reflects correlations in the raters' perceptions of the different dimensions that are evaluated in analytic scoring.

Online Team Training, Rater Monitoring Systems, and Rater Accuracy

Two studies using operational rater-monitoring data from high stakes examinations in England are contrasted. Rater accuracy was analyzed using cross-classified hierarchical linear modeling.

The "Third Rater Fallacy" in Essay Rating: An Empirical Test

This paper seeks to test empirically the benefit of using a third rater in cases of disagreement between two raters. It is shown that empirical data are in agreement with classical test theory (CTT), viewing each rating as a sum of true score and an error component. The data also corroborate results that were obtained in computer simulations based on CTT.

Saturday April 18, 2015
2:15 PM - 3:45 PM, Grand Ballroom, 7th Floor, Coordinated Session, H4

Evaluating and Improving Methods for Student Growth Percentile Estimation

Session Chair: J.R. Lockwood, Educational Testing Service
Session Discussant: Derek Briggs, University of Colorado Boulder

Presenters: J.R. Lockwood, Daniel McCaffrey, Katherine Castellano, Educational Testing Service; Elias Walsh, Mathematica Policy Research; and Harold Doran, American Institute for Research

Student growth percentile (SGP) and derived measures are being used across the country to make impactful decisions for students, teachers, and school leaders. This coordinated session will present cutting-edge research on measurement properties of SGP estimators as well as alternative statistical approaches to their computation.

How Does Teacher Value Added Compare to Median Growth Percentiles?
Elias Walsh & Eric Isenberg, Mathematica Policy Research

This paper compares teacher value-added estimates to median growth percentiles using real data. It finds that the measures can differ systematically in ways that relate to the background characteristics of the students taught by different teachers. Differences may due to how the SGP estimation process adjusts for prior achievement.

Alternative Approaches for Computing SGP with Mismeasured Variables
Harold Doran, American Institute for Research

This paper introduces an alternative method for SGP estimation, currently being used by New York State that accounts for measurement error in test scores and requires fewer model parameters. We demonstrate via simulation its ability to reduce mean squared error relative to the current approach.

Improved Statistical Frameworks for Student Growth Percentile Estimation
J.R. Lockwood and Katherine E. Castellano, Educational Testing Service

This paper provides two alternative statistical frameworks for improving the estimation of SGP. The first is to estimate SGP directly by modeling conditional cumulative distribution functions rather than indirectly through quantile regressions. The second is to estimate SGP directly from longitudinal item-level data using multidimensional item response theory models.

Measurement Error Bias for Student- and Group-Level Student Growth Percentiles
Daniel F. McCaffrey and Katherine E. Castellano, Educational Testing Service

This paper demonstrates the effects of measurement error in both the past and current test scores on standard and alternative SGP and MGP estimators. It develops analytical results on the mean and standard deviation of the estimators under normality assumptions. It discusses the bias-variance tradeoffs of the estimators and the corresponding implications at the student- and aggregate-level.

Saturday, April 18, 2015
2:15 p.m.-3:45 p.m., Renaissance Ballroom, 5th Floor, Invited Session, H5

The Importance of Instructional Sensitivity: A Colloquy Among Combatants

Session Chair: Ronald K. Hambleton, University of Massachusetts Amherst

W. James Popham, University of California Los Angeles

When, several decades ago, it was recognized that many of our nation's most widely used educational tests contained meaningfully biased items, the measurement community tackled this problem and, consequently, dramatically reduced such bias in America's educational tests. Currently, we find many high-stakes tests being used to evaluate the instructional success of both schools and teachers. Yet, this is occurring despite the complete absence of evidence supporting such an evaluative application. If today, members of the educational measurement community fail to resolve this problem, we will be committing a psychometric sin so serious that it could cripple our field forever.

Neal Kingston, University of Kansas

Little evidence exists that test items can differentiate between students who were and were not taught the content of the items. If such a gross differentiation is not possible, how can we justify finer differentiation, such as attributing differences in test results to the quality of teachers? If we want to use test results as a primary basis for school or teacher accountability we need to understand why so few items show evidence of sensitivity to instruction and use this information to create tests that better serve our policy goals.

Denny Way, Pearson and John Fremer, Caveon

The position "against" instructional sensitivity is expressed in the context of large-scale, standard-based assessments. No consistent empirical procedures for detecting the instructional sensitivity of questions for these assessments have been established or are even promising, although the idea is conceptually attractive. Furthermore, as best as we can tell, the items most likely to be impacted by instruction measure lower order skills - such as dates, facts, special vocabulary, and simple relationships. The things harder to teach require understanding, seeing relationships, recognizing patterns, modeling, etc., and the simple notion of instructional sensitivity does not easily fit with items measuring these more complex skills. We argue that the best evidence of instructional sensitivity is found in documentation related to what is supposed to be taught in the classroom, what is actually taught in the classroom, how well tests and items align with what is taught, and the quality of their psychometric properties.

Saturday April 18, 2015
2:15 PM - 3:45 PM, King Arthur, 3rd Floor, Coordinated Session, H6

Smarter Balanced Automated Scoring Research: Results and Insights

Session Chair: Vincent Keiftenbeld, McGraw-Hill Education CTB
Session Discussant: David Williamson, Educational Testing Service

The Smarter Balanced Assessment Consortium is a state-led consortium working to develop next-generation assessments that accurately measure student progress toward college- and career-readiness. In 2013 and 2014, the consortium conducted research to further the state-of-the-art in automated scoring. This session reports on the results of from the Pilot and Field Test. The findings are relevant for researchers in automated scoring, educational policy makers, and other stakeholders in large-scale assessments of college and career readiness.

Performance and Evaluation of Automated Scoring Models
Vincent Kieftenbeld, McGraw-Hill Education CTB

How well do current state-of-the-art automated scoring system rate student responses on constructed-response items compared to the gold standard of adjudicated human scores? This presentation reports on the performance of automated scoring models trained by nine different vendors for equation, short-text (English language arts and mathematics), and essay items during the Pilot and Field Test. Performance was evaluated using several criteria, including exact agreement rates, quadratic weighted kappa, and standardized mean differences.

Targeting Responses for Human Review
Frank Rijmen, McGraw-Hill Education CTB

After training and validating automated scoring systems, a large-scale assessment program can deploy the resulting models in several scoring scenarios. These scoring scenarios range from fully automated operational scoring to combinations of automated and human scoring. Although the former are efficient (after training), the latter offer several benefits related to score quality. Some student responses may require human review, for example, when responses are unlike the responses using in training to the extent that this may affect scoring accuracy. This presentation reports on three different methods that were used to identify responses likely to require human review.

Reducing Development Costs of Automated Scoring
Claudia Leacock, McGraw-Hill Education CTB

Although automated scoring may be efficient (in terms of time as well as costs) once scoring starts, the initial development costs can be high. This presentation focuses on two methods to potentially reduce the development costs. The first method considers item characteristics that can predict whether a short-text, constructed-response, item can be scored automatically. Several item characteristics were found to correlate with subsequent performance of the automated scoring systems. The second method reduced development costs by training a single automated scoring system to score grammar and writing conventions across several essay prompts, rather than training a scoring model for each item separately.

The Role of Automated Scoring in Smarter Balanced
Joe Willhoft, Smarter Balanced Assessment Consortium
Discussant: David Williamson, Educational Testing Service

Where the first three presentations highlight various technical aspects of the operational use of automated scoring, this presentation focuses on synthesizing insights gained from the Smarter Balanced automated scoring research from the perspective of educational policy makers and other stakeholders in college and career readiness assessments. Looking back at the Pilot and Field Test, it articulates policy and implementation issues that need to consider when incorporating automated scoring in their assessments. Topics relevant to automated scoring of the next-generation assessments are considered, including the relationship between college and career readiness standards on the one hand and automated scoring on the other, and policies and best practices related to quality assurance and accountability. Looking forward, this presentation discusses the future of automated scoring in the context of the Smarter Balanced assessments specifically and college and career readiness assessments generally.

Saturday April 18, 2015
2:15 PM - 3:45 PM, Seville Ballroom East, Lobby Level,
Coordinated Session, H7

Third Grade Reading Proficiency: Two Large-Scale Longitudinal Studies

Session Chair: James McBride, Renaissance Learning, Inc.
Session Discussant: John Sabatini, Educational Testing Service

Presenters: James McBride, James Olsen, Philip Giesy, Renaissance Learning, Inc., and Mike Beck, Beta, Inc.

This session explores aspects of early literacy assessment for predicting later third-grade reading proficiency. Papers address the design and development of the early literacy and reading assessments, a national proficiency standard setting for reading assessment, and two longitudinal research studies, one at a macro-level and the other at a micro-level.

Overview: Adaptive Testing of Pre-Literacy Skills and Reading Achievement
James R. McBride, Renaissance Learning, Inc.

This paper introduces the early literacy and reading assessments used in the two analytic studies. These assessments measure student position in a learning progression spanning from pre-literacy through reading comprehension of complex texts. The paper describes the common scale development and illustrates its uses in validating the learning progression.

Setting National Performance Standards for Reading Proficiency in Primary Grades
Mike Beck, BETA Inc.

A modified Bookmark standard setting was used to establish performance levels on a widely used assessment. "Focused" ordered-item booklets were employed, involving unique sets of items clustered around the panel's initial recommendations. Data concerning the "Focused Bookmark" approach for setting standards for adaptive tests with sizeable item pools are presented.

Predicting Early Reading Proficiency Using Pre-School Early Literacy Scores
James B. Olsen and James R. McBride, Renaissance Learning, Inc.

This paper presents a longitudinal study of prediction of third grade reading proficiency levels and scaled scores from Pre-K and Kindergarten early literacy assessments. The study examines the predictive relationship of early pre-literacy status measures and same students' reading proficiency measures, four years later, at the end of third grade.

Which Early Literacy Skills Best Predict Grade 3 Reading Proficiency?
Philip J. Giesy, Renaissance Learning, Inc.

The paper describes a longitudinal study measuring the correlation between students' success with each of more than 100 specific early literacy skills (Pre-Kindergarten through 1st grade) with later reading success at grade 3. A Predictive Index was developed that ranks the early literacy skills in terms of predictive value.

Saturday April 18, 2015
2:15 PM - 3:45 PM, Seville Ballroom West, Lobby Level,
Coordinated Session, H8

Feasibility of Various Cut Score Moderation Methods

Session Chair: Priya Kannan, Educational Testing Service
Session Discussant: Gregory J. Cizek, University of North Carolina Chapel Hill

This coordinated session explores the idea of cut score articulation /moderation for related tests. Several methodological solutions to articulate or moderate cut scores have been offered. This session explores the feasibility of a few cut score moderation methods (e.g., logistic regressions) and offers other novel methodological solutions (e.g., policy linking).

Impact of Grade-Level Correlation on Classification Consistency of Articulated Cut-Scores
Priya Kannan and Adrienne Sgammato, Educational Testing Service

This simulation study evaluated the effect of varying the underlying grade-level correlations on the bias and classification consistency of cut scores articulated using two methods (i.e., logistic regression and equipercentile smoothing). Results suggest that small underlying correlations result in significant bias and lower classification consistency for the logistic regression method.

Linking Considerations for Logistic Regression-Based Articulation Methods
Adrienne Sgammato and Priya Kannan, Educational Testing Service

Bias of the logistic regression method was evaluated in recovering 'true' cut-scores (established assuming perfect correlation between grades) when the underlying correlations across grade levels varied. Alternative grade-level linking procedures were evaluated. Results indicate that all linking methods resulted in biased cut scores; some were less biased than others.

Developing College and Career Readiness Cut Scores in One State
Shiqi Hao, Michigan Department of Education and Adam E. Wyse, The American Registry of Radiologic Technologists

This study compares the use of three moderation methods (logistic regression, signal detection theory, and equipercentile cohort matching) to create college and career readiness cut scores for mathematics, reading, science, and social studies in a large Midwestern state. Results suggested some differences between methods across grades and subjects.

Policy Linking as Cut Score Moderation: Considerations for Practice
Chad W. Buckendahl and Brett P. Foley, Alpine Testing Solutions

Vertical moderation of cut scores has been used to facilitate coherence across grade levels. Moderation can also occur across assessments designed to measure the same construct through policy linking. This paper suggests a framework, illustrating it with a combination of policy descriptors and supplemental cut score information.

Saturday April 18, 2015
2:15 PM - 3:45 PM, Toledo, 5th Floor, Coordinated Session, H9

Research and Development on Assessment and Accountability for Special Education

Session Chair: Joseph Stevens, University of Oregon
Session Discussant: Barbara Plake, University of Nebraska-Lincoln, Emeritus

Presenters: Ann Schulte, Arizona State University; Gerald Tindal, Joseph Nese, University of Oregon; Stephen Elliott, Alexander Kurz, Arizona State University; and Joseph Stevens, University of Oregon

This session provides information on the National Center for Assessment and Accountability for Special Education which studies the achievement growth of students with and without disabilities. The session purpose is to provide an overview of the center, follow-up on presentations at two previous NCME conferences, and present four research papers.

Does One Size Fit All? Reading Achievement Growth for Students With and Without Disabilities
Ann C. Schulte, Arizona State University and Joseph Stevens, University of Oregon

Reading growth across five years in a state-wide cohort (N = 94,650) was contrasted for students in general education and within seven disability classifications. Growth was curvilinear, intercept was negatively correlated with slope, and students with disabilities generally differed from students in general education in both intercept and slope.

Modeling Growth for NCLB Subgroups: Effects of Time-Varying Disability Classification
Joseph F.T. Nese, Gerald Tindal, Joseph J. Stevens, University of Oregon; Ann C. Schulte, and Stephen N. Elliott, Arizona State University

The purpose of this paper is to compare trajectory estimates of quadratic growth models for time-varying versus fixed special education exceptionality categories. We use statewide achievement test data and analyze the bias of model parameters by exceptionality modeling decision (fixed or varying) for two cohorts, Grades 3-7.

Alternative Methods for Computing Growth Norms
Joseph J. Stevens, Joseph F.T. Nese, and Gerald Tindal, University or Oregon

This paper describes alternative methods for creating growth norms to describe and benchmark academic achievement. Medical norms, student growth percentiles, and multilevel model norms are described, along with their advantages and disadvantages. Analyses use statewide accountability data from Oregon and an interim assessment administered in a large Arizona school district.

Predicting End-of-Year Mathematics Achievement of Students With and Without Disabilities: The Role of Opportunity to Learn and CBM Measures

Stephen N. Elliott, Alexander Kurz, Arizona State University; Gerald Tindal, University of Oregon, and Nedim Yel, Arizona State University

We examined how teachers' instructional processes and progress monitoring measures predicted students' end-of-year achievement in mathematics on their statewide achievement tests. The results supported the prediction that opportunity to learn (OTL) indices and CBM scores accounted for substantial variance in elementary and secondary students' end-of-year mathematics performance.

Saturday April 18, 2015
2:15 PM - 3:45 PM, Valencia, Lobby Level, Paper Session, H10

Smoothing in Equating

Session Chair: Ying Chen, University of Notre Dame
Session Discussant: Amy Hendrickson, College Board

Why Observed-Score Equating Transformations Look Like This
Jiahui Zhang, Michigan State University and Wei Tian, Beijing Normal University

Equipercentile equating is a general method of observed-score equating. The shapes and locations of equipercentile equating transformations haven't received adequate attention. The analysis of equipercentile equating transformations turns out to provide evidence against traditional equipercentile equating and also establish the necessity of conducting local equating.

A Comprehensive Comparison of Smoothing Methods for the CINEG Design
Han Yi Kim, Measured Progress; Won-Chan Lee, and Walter Vispoel, University of Iowa

Relative performances of smoothing methods were compared under the common item nonequivalent groups (CINEG) design with 192 simulated testing conditions. Results showed that log-linear presmoothing produced smaller total error under more testing conditions than did cubic spline postsmoothing. This study will produce general guidelines into the selection of smoothing procedures.

Selection Strategies for Loglinear Smoothing Models in a NEAT Design
Weldon Smith and Anthony Albano, University of Nebraska-Lincoln, Lincoln, NE

Prior to conducting equipercentile equating, various model comparison strategies are used to select a loglinear model with an appropriate amount of distribution smoothing. This study aims to investigate the relationships between model selection, sample size, test length and the resulting introduction of bias and reduction of standard errors in equating.

Issues Regarding Structural Zeros in Bivariate Log-Linear Presmoothing
Hyung Jin Kim, Robert Brennan, and Won-Chan Lee, The University of Iowa, Iowa City, IA

Structural zeros are cells with zero probabilities of observing pairs of scores in a bivariate distribution. When presmoothing assigns positive probabilities to structural zeros, they become a source of bias in equating. This study examines different approaches to handle structural zeros and investigates how equating results compare for different approaches.

Saturday April 18, 2015
4:05 PM - 5:05 PM, Camelot, 3rd Floor, Electronic Board Session, Paper Session, I1

Electronic Board #1
Performance of Relative Fit Indices: A Comparison Across Model Types
Sedat Sen, Harran University and Laine Bradshaw, University of Georgia

Model-data fit plays an important role in making valid model-based interpretations. The performances of three relative fit indices (AIC, BIC, and SABIC) were examined under various simulated conditions. Conditions involved comparisons across cognitive diagnosis models (CDMs) and item response theory (IRT) models. Suggestions were made for researchers based on results.

Electronic Board #2
Anchor Purification: Tagging IPD Flags to Construct-Relevant Variance
Alvaro Arce-Ferrer, Pearson and Avi Allalouf, NITE

The paper studies the behavior of a new approach to inform item retention-dropping decisions when linking similar tests with different content standards with the common item non-equivalent groups design. The process provides a valid way to gauge the intrinsic value of the anchor set. The paper discusses recommendations and research.

Electronic Board #3
Standard Errors for National Trends in International Large-Scale Assessments
Karoline Sachse and Nicole Haag, Institute for Educational Quality Improvement, Humboldt-University of Berlin, Berlin, Germany

We compared different approaches to standard error computations for national trends in international comparative large-scale assessments using simulated data. Specifically, we investigated how standard errors can be estimated most accurately if a latent ability shift, cross-national DIF, and item parameter drift are present. Results and practical implications are discussed.

Electronic Board #4
Evaluating the Psychometric Properties of Generated Test Items
Mark Gierl, Hollis Lai, University of Alberta; Andre-Philippe Boulais, and Andre De Champlain, Medical Council of Canada

The purpose of this study is to present the first empirical results describing the psychometric quality of generated test items administered as part of an operational test administration. The item analysis results for both the correct option and the distractors are reported using measures from classical and modern test theory.

Electronic Board #5
Effect of Difference in Reliability Between Tests on Linking
Nooree Huh, ACT, Inc., Iowa City

In this study, the relationship between differences in test score reliabilities in linked tests and the final conversion tables is examined based on simulated data in a single group design. In addition, a decision consistency rate based on a raw cut score is examined.

Electronic Board #6
Modeling Student Test-Taking Motivation in the Context of CAT
Steven Wise, Northwest Evaluation Association and Gage Kingsbury, Psychometric Consultant

When a CAT is administered, non-effortful test taking may result in both negatively-biased achievement estimation and mis-targeting of item difficulty. This study introduces a CAT procedure that may mitigate these problems. This represents a viable alternative to simply invalidating scores from non-effortful test events.

Electronic Board #7
Applying Pre-Equating on Exams With Small Sample Size
MinJeong Shin, Chi-Yu Huang, and Meichu Fan, ACT, Inc.

This study investigated the impact of small sample size, and the applicability of converting classical item statistics to IRT item parameters, on pre-equating. Initial results of this study show that IRT parameters converted from classical statistics produced stable item parameter recovery results as well as acceptable equating conversions.

Electronic Board #8
An Investigation of IRT Reliability for Technology Enhanced Items
Dong-In Kim, Wen-Ching Lee, Litong Zhang, and Sara Kendall, and McGraw-Hill Education CTB

TE items are considered to have the characteristics of SR item's scoring efficiency and CR item's desirable cognitive complexity and Depth of Knowledge. Several different types of TE items will be analyzed with various combinations of scoring methods and IRT scaling models using large-scale high school Algebra tests.

Electronic Board #9
Item Parameter Estimation Methods: Achievement and System Ranking Stability
Leslie Rutkowski, David Rutkowski, and Yan Zhou, Indiana University, Bloomington, IN

Using a simulation study based on empirical PISA 2009 results, we evaluate the performance of three methods of item parameter estimation on proficiency mean recovery and system rankings. We found better true mean coverage under the method currently in use and unstable rankings for middle performers under two estimation approaches.

Electronic Board #10
Constructing Balanced Incomplete Block Designs Tests: Ant Colony Optimization
Pei-Hua Chen, National Chiao Tung University and Wan-Yu Tsai, Florida State University

Three content balancing methods were proposed and incorporated with the Ant Colony Optimization approach for test assembly. Thirteen booklets with thirteen five-item blocks were constructed from a 292-item bank. Preliminary results show that the fixed content ratio method performs well in terms of computation time and measurement precision.

Electronic Board #11
Assessing the Item Fit in Computerized Adaptive Tests
Yuan Hong, Tao Jiang, and Stephan Ahadi, American Institutes for Research, Washington, DC

As state start using the immediate score reporting offered by the online assessments, it becomes crucial to maintain a calibrated item pool with pre-equated parameters. Stone's item fit statistic is used to flag the non-fit items and is generalized to be able to use both MLE and EAP ability estimates.

Electronic Board #12

Mixed-format Multistage Tests Under the 3PL Testlet Response Theory Model

Ian Hembry, Amplify Learning and Barbara Dodd, The University of Texas at Austin

The study examined operational characteristics of multistage test designs under a three-parameter logistic testlet response theory model. Simulation conditions included four panel designs, two test lengths, three routing procedures, and three local item dependence conditions. No bias was detected, but measurement precision was higher for the two-stage panel designs.

Electronic Board #13

Sensitivity of Fit Indices to Q-Matrix Misspecification in the CRE-LLTM

Chunhua Cao, Yi-Hsin Chen, Isaac Li, and Yan Wang, University of South Florida, Tampa

This simulation study examined the sensitivity of commonly used fit indices to Q-matrix misspecification in cross random effects linear logistic test model (CRE-LLTM). The fit indices examined in this study include -2LL, AIC, AICc, BIC, and HQIC. The impact of the design factors on the fit indices was also investigated.

Electronic Board #14

Meta-Analysis to Assess Generic Shift of Exam Populations

Hendrik Straat and Marieke van Onna, Cito, Arnhem, Netherlands

Due to changes in the Dutch exam system, an alternative had to be found for linking under the randomly equivalent populations assumption. A NEAT design on some subjects combined with a Fisher's method for meta-analysis was applied to evaluate if exam candidates generically performed differently in 2014 compared to 2011.

Electronic Board #15

Online Calibration for a Joint Model of Responses and Response Times in CAT

Hyeon-Ah Kang, University of Illinois at Urbana-Champaign; Yi Zheng, Arizona State University Tempe; and Hua-Hua Chang, University of Illinois at Urbana-Champaign

In CAT, a sparse response matrix constitutes a challenge to accurately calibrating pretest item parameters. In this study, we show how much statistical improvement can be obtained in online calibration by capitalizing on response times (RTs), and propose adaptive online calibration methods that incorporate both item responses and RTs.

Electronic Board #16

An Iterative Method of Empirically-Based Q-Matrix Validation

Ragip Terzi and Jimmy de la Torre, Rutgers, The State University of New Jersey, New Brunswick, NJ

In cognitive diagnosis modeling, constructing a Q-matrix can be subjective, resulting in serious validity concerns. Misspecifications in the Q-matrix severely affect parameter estimation, and ultimately attribute-classification. To address this issue, we propose a Q-matrix validation method based on an existing procedure, where iteration and various cut-off points are added.

Electronic Board #17

Application and Justification of Vertical Comparison Instead of Vertical Linking

Anton Béguin and Saskia Wools, Cito Institute for Educational Measurement, Arnhem, Netherlands

In this paper it is shown that vertical comparison using reference sets is less restrictive than current vertical linking procedures. Results are given of a study that uses vertical comparison to set equivalent performance standards in vertically different populations and a small simulation study shows the effectiveness of vertical comparison.

Electronic Board #18
IRT Equating for Test With Seasonality
Yanming Jiang and Yuming Liu, Educational Testing Service, Princeton

We examine effects of seasonality on equating when multiple equatings are performed. Varying levels of seasonality are considered, which reflect the extent of group ability differences across administrations. We investigate how the magnitude and patterns of seasonality affect equating results, and seek an appropriate equating method for tests exhibiting seasonality.

Electronic Board #19
Relationships of Growth Measures From Different Plausible Vertical Scales
Dongmei Li, ACT, Inc., Iowa City

There are many different but maybe equally defensible ways to construct a vertical scale. How would rank orders of student growth change across the many plausible scales? This study investigates the relationships of growth measures from various potential scales, when growth is measured using either simple or residual gain scores.

Electronic Board #20
Allowing Unrestricted Answer Changing Through Computerized Adaptive Testing With Salt
Zhongmin Cui, Chunyan Liu, Yong He, and Hanwei Chen, ACT, Inc., Iowa City

We proposed and evaluated a new computerized adaptive testing procedure to provide examinees unrestricted opportunities to review and make answer changes to any test item at any time before submitting the whole test. The new procedure was shown to be efficient in estimating abilities while being robust to cheating strategies.

Electronic Board #21
Modeling and Evaluating Generated Items Under Common Core Standards
Hollis Lai, Mark Gierl, University of Alberta; and Jim Hogan, ACT

Model-based item generation can produce large numbers of test items. An item model contains specification of required skills and knowledge to produce the correct response. By aligning common core standards within the modeling process, we propose a method to generate items with the required standards and evaluate their alignment.

Electronic Board #22
Using Response Time for Scoring With Applications to Multistage Testing
Usama Ali, Educational Testing Service and Peter van Rijn, Educational Testing Service Global

We investigate the use if response time (RT) in adaptive testing. Combining accuracy and speed in scoring might improve the quality of measurement by (1) enhancing the item selection algorithm in item-level adaptive tests and (2) creating easy routing blocks in multistage tests without loosing precision.

Electronic Board #23
Effect of Population Changes on Equating: Analysis of Repeater Data
Jennifer Dunn, Han Yi Kim, Louis Roussos, Wonsuk Kim, and Andrew Martin, Measured Progress, Dover, NH

Equating results are often influenced by changes in the student population. The purpose of this study is to examine the effects of the inclusion or exclusion of repeaters in test equating and student classification for the CINEG design, and to evaluate the practical significance of those effects.

Electronic Board #24

The Effects of Items With Undetectable DIF on Equating

Xiaoran Li and Jane Rogers, University of Connecticut, Storrs, CT

Anchor items with Differential Item Functioning (DIF) could cause substantial problems in equating. However, the DIF items may not be detected due to power constraints, and are retained for equating. This study defines a level of undetectable DIF, and examines the accuracy of equating with undetectable DIF items.

Electronic Board #25

An IRT-Based Method for Detecting Compromised Items in CAT

Jinming Zhang, University of Illinois at Urbana-Champaign and Jie Li, McGraw-Hill Education CTB

An IRT-based sequential procedure is developed to identify compromised items in CAT by examining whether the statistical characteristics of individual items have changed significantly during CAT administration. Simulation studies show that it can control the rate of type I errors and has a very low rate of type II errors.

Saturday April 18, 2015
4:05 PM - 6:05 PM, St. Clair, Upper 5th Floor, Coordinated Session, I2

Evaluating Scoring Issues for Innovative and Technology Enhanced Items

Session Chair: Joseph Betts, Pearson VUE
Session Discussant: Joshua Goodman, Pacific Metrics

This session will focus on scoring innovative, technology-enhanced items (TEI). The papers will provide an overview of the application of an evidence-centered design as a general framework for scoring items, specific investigations of item types appropriate for a number of testing programs, and guidance on sample size requirements for calibration.

Constructing a Framework for Scoring Innovative Test Items
Xiao Luo, National Council of State Boards of Nursing; Kirk Becker, Karen Sutherland, Pearson VUE; and John De Jong, Pearson

An evidence-based scoring method for innovative item types is described and illustrated with several practical examples in the context of licensure testing. This method describes how to extract and score meaningful evidentiary objects from responses.

Scoring Options for Ordered Response Items
Kirk Becker, Pearson VUE, Hong Qian, National Council of State Boards of Nursing, and Karen Sutherland, Pearson VUE

This research explores the psychometric characteristics of scores for ordered list items which allow for ordering of certain steps relative to "milestone" steps.

Analyzing Multiple Response Data Through a Signal-detection Framework
William Muntean, Joseph Betts, Pearson VUE; and Ada Woo, National Council of State Boards of Nursing

Multiple response formats are well suited for assessing the construct of cue recognition, the discrimination between information that is relevant or irrelevant to a decision. One useful approach to analyzing cue recognition data is through a signal-detection measurement model. This proposal investigates the ideal number of cues required to achieve reasonable measurements.

Investigating Sample Size Requirements for the Partial Credit Model
Joseph Betts, Pearson VUE, Doyoung Kim, National Council of State Boards of Nursing, and William Muntean, Pearson VUE

This presentation will report on the results of a simulation study evaluating sample size requirements for calibrating items using the partial credit model. Factors manipulated were number of thresholds per item, range of distances between thresholds, overall sample size, and number of response categories. Model parameter recovery will be evaluated.

Saturday April 18, 2015
4:05 PM - 6:05 PM, Exchange, 11th Floor, Coordinated Session, I3

Psychometrics in a Learning Maps Environment

Session Chair: Amy Clark, University of Kansas
Session Discussant: Russell Almond, Florida State University

Session Chair: Amy Clark, University of Kansas Session
Discussant: Russell Almond, Florida State University
Presenters: Amy Clark, University of Kansas; Jonathan Templin, Neal Kingston, University of Kansas; and Laine Bradshaw, University of Georgia

The session explores approaches to psychometrics in a dynamic assessment environment based on learning maps. Four papers will be presented, focusing on 1) the differences between node mastery and traditional scale scores, 2) item analysis and item selection, 3) reliability, and 4) the nature of standards setting, growth, and equating.

Saturday April 18, 2015
4:05 PM - 6:05 PM, Seville Ballroom East, Lobby Level, Invited Session, I4

A Dialogue for Addressing Measurement and Data Gaps in Education

Session Co-Chairs: Joshua Marland and Lisa Keller, University of Massachusetts Amherst

As education data proliferate, educators are expected to make sense of the many measures that could be potentially used to enhance learning and improve practice. At the same time, educators are criticized for not using all available data – all without consideration for the barriers, gaps and tensions that may exist related to use. Teachers, teacher educators and psychometricians discuss the ways in which they currently use or provide data that are intended to enhance student learning and improve practice, as well as barriers to use and gaps in existing measures. Panelists will also discuss future directions for data use in their respective roles with a specific eye toward current best practices.

Panelists:
Drey Martone, College of Saint Rose
Charlie DePascale, National Center for the Improvement of Educational Assessment
Kristen Huff, Regents Research Fund

Teachers and principals will discuss the types of data they use on a regular basis, the challenges they face to using data, and where they would like the measurement field to go in providing valuable information to them.

Drey Martone will discuss the data processes she teaches to future educators, some of the necessary conditions for success, challenges to implementing good data practices in the classroom, and the existing gaps in available measures to educators.

Charlie DePascale and Kristen Huff will talk about psychometricians' roles in ensuring high-quality measures, competing measurement priorities for, and barriers to, providing educators with more useful information,

Saturday April 18, 2015
4:05 PM - 6:05 PM, King Arthur, 3rd Floor, Coordinated Session, I5

Surf and Turf Summative Assessment: States Combining Efficiencies With Customization

Session Chair: Marianne Perie, Center for Educational Testing and Evaluation
Session Discussant: Michael Kolen, University of Iowa

Presenters: Marianne Perie, Center for Educational Testing & Evaluation; Erik McCormick, Alaska Department of Education and Early Development; Scott Smith, Kansas State Department of Education; Gail Tiemann, and Laura Kramer, Center for Educational Testing & Evaluation

Apart from the main consortia, two states and a university collaborated on a summative assessment that maximizes efficiencies while also allowing for customization within each state. This session will describe the policy decisions, the multi-stage assessment design including innovative items, and balancing scales, performance level descriptors, and cut scores.

Saturday April 18, 2015
4:05 PM - 6:05 PM, Renaissance Ballroom, 5th Floor, Invited Session, I6

Standard Setting in the Common Core World: PARCC and SBAC Experiences

Session Chair: Leslie Keng, Pearson
Session Discussant: Gregory J. Cizek, University of North Carolina Chapel Hill

Policy Considerations in PARCC Standard Setting
Enis Dogan and Stephanie Snyder, PARCC Inc.

PARCC standard setting panels will be convened in the summer 2015. In this presentation we will discuss how the performance level descriptors were developed, the role of empirical data in standard setting, considerations in selection of panelists, and the process that will be followed in approving the cut scores.

PARCC Standard Setting Methodology
Julie Miles, Pearson VUE

An overview of the critical elements of the Evidence Based Standard Setting process which is used to integrate empirical data from systematic research and content expert judgment in setting performance standards. The seven steps of the EBSS method used in supporting the PARCC standard setting process will be highlighted.

Policy Considerations in SBAC Standard Setting
Joe Willhoft, SBAC

Policy considerations for standard setting are usually entangled with communications considerations. This presentation will focus on three such considerations: the identification and selection of scores to use for impact data, the management of a diverse set of policy-level stakeholders, and clarifying the role of performance levels in reporting.

Setting Cut Scores on Smarter Balanced Assessments: A Ground-Level View
Michael Bunch, Measurement Inc.

This paper focuses on the preparations for and execution of a Bookmark procedure for an on-site/online achievement level setting involving over 3,000 panelists in the setting of cut scores for 14 Smarter Balanced tests in the fall of 2014. Recruiting, communications, and logistics are also discussed.

Saturday April 18, 2015
4:05 PM - 6:05 PM, Adler, 2nd Floor, Paper Session, I7

Person Fit and Aberrant Responses

Session Chair: Su Baldwin, National Board of Medical Examiners
Session Discussant: Jane Rogers, University of Connecticut

Effect of Successions of Same Responses on Answer Copying Detection
Hongling Wang and Chi-Yu Huang, ACT, Inc, Iowa City, IA

This study explores the effect of successions of same responses on answer copying detection by statistical indices of response similarity. Successions of same responses may lead to increased type I and type II errors. The results will provide test practitioners a guideline for dealing with this issue.

The Use of Person Fit Indices in Multidimensional Structure Data
Yang Lu and Yu Fang, ACT Inc., Iowa City, IA

The purpose of this study is to evaluate the effectiveness of three person fit indices (l_z, l_zm and H^T) to detect the four aberrant response patterns on the multidimensional structure data. The results will show the impacts of estimation model on the performance of person fit indices.

Robust Estimation of Latent Abilities for Speeded Test-Takers
Chien-Lin Yang, Haiqin Chen, American Dental Association; and Paul De Boeck, The Ohio State University

To improve the ability estimation for speeded test-takers, robust estimation methods are used to downweight the influence of speeded test responses. The performance of maximum likelihood and robust estimators are compared through a simulation study and an empirical data set with identified speeded test-takers.

Examining Erasure Behaviors in Large-Scale Assessment
Elizabeth Ayers, American Institutes for Research and Yoonjeong Kang, University of Maryland

Data from a 5th grade reading assessment is used to examine whether aberrant erasure behaviors are associated with irregular item responses from cheating. Erasure analyses using person-fit statistics and modified test-retest methods along with erasure indices show that aberrant erasure behaviors are possibly related to aberrant response patterns from cheating.

Ability Estimation in the Presence of Aberrant Responses
Hua Wei and Tian Song, Pearson, Cincinnati

This study compares two ability estimation approaches in the presence of aberrant responses in terms of recovery of ability parameters and overall model fit. Findings of the study have practical implications for model selection and ability estimation for response data that are contaminated with guessing and carelessness.

Saturday April 18, 2015
4:05 PM - 6:05 PM, Seville Ballroom West, Lobby Level, Paper Session, I8

DIF: Sample Size, Effect Size, Power

Session Chair: Lee Lafond, Measured Progress
Session Discussant: Ahmet Turhan, Pearson

Detecting Differential Item Functioning of Tests for Special Populations
Kwang-lee Chu, Pearson, Pei-ying Lin, University of Saskatchewan, Saskatoon, and Marc Johnson, Pearson

The impact of disabilities on DIF analysis is examined through three DIF models. Empirical data is examined first and then used for simulations investigating accuracy of DIF, effect of sample size, and impact of matching group slicing within the three DIF models.

DIF Analyses Between Groups When Size and Proficiency Distributions Differ
Lynne Hollingshead, University of Toronto, Newmarket, Canada

This study investigates the use of a bootstrap method to improve contingency table DIF analyses of groups with differing size and proficiency distributions, contrasting this approach with the current recommendations in the literature.

Statistical Power for Assessing Measurement Invariance in Latent Profile Analysis
Margarita Olivera Aguilar, Educational Testing Service Global and Samuel Rikoon, Educational Testing Service

The study of invariance in latent profile analysis (LPA) indicates if the number and nature of the profiles differ across known subgroups (e.g. gender). In this simulation study we will examine Type I error rates and statistical power of five fit statistics for detecting violations of invariance in LPA.

Comparison of Different Methods to Enhance Small Sample DIF Estimation
Xiuyuan Zhang, Anita Rawls, Weiwei Cui, and Amy Hendrickson, The College Board

The Mantel-Haenszel (MH) procedure is widely used in operational contexts for identifying items with differential item functioning (DIF). Motivated by a realistic condition, the present study intended to investigate the possibility of using distribution smoothing or thick matching of criterion scores in MH DIF estimation with small samples.

Power and Sample Size Formulas for Mantel-Haenszel DIF Test
Zhushan Li, Boston College, Chestnut Hill

A power formula for the Mantel-Haenszel test for differential item functioning (DIF) is derived. It provides a means for calculating the sample size in planning DIF studies with MH test. Factors influencing the power are discussed. The correctness of the power formula is confirmed by simulation studies.

Longitudinal Measurement Invariance When Both People and Items Change
Ronli Diakow, New York University, New York, NY

This paper explicates the link between identification constraints, measurement invariance, and interpretation in models where both person and item parameters can vary among groups over time. Statistical properties of different constraints are explored via simulations; the influence of constraints on interpretation is addressed using empirical data from an efficacy study.

Saturday April 18, 2015
4:05 PM - 6:05 PM, Toledo, 5th Floor, Paper Session, I9

Extraneous Factors Affecting Test Behaviors

Session Chair: Tracey Hembry, Alpine Testing
Session Discussant: Katrina Crotts Roohr, Educational Testing Service

The Effect of Option Homogeneity in Multiple Choice Items
Gregory M. Applegate, Karen A. Sutherland, Pearson; and Xiao Luo, National Council of State Boards of Nursing

The effect of option homogeneity on item parameters in multiple-choice items was tested empirically. Similarity of options to the key was determined using subject matter experts and a natural language processing algorithm. While previous research suggests option homogeneity would affect item parameters, our findings contradict that literature.

Examining Item Order Effects on Test Scores in Online Testing
Jungnam Kim, National Board of Chiropractic Examiners; Furong Gao, Pacific Metric; Ping Wan, Sandra McGuiire, and Dong-In Kim, McGraw-Hill Education CTB

In online testing, alternate test forms of the same items, in different item locations, has been employed as a way to improve test security. The item order effect in these alternate test forms of a large scale assessment is investigated at test and item level using IRT and G-theory.

The Time on Task Effect in Digital Reading Items
Johannes Naumann, Goethe-University and Frank Goldhammer, German Institute for International Educational Research

Using PISA 2009 Digital Reading Assessment data (N=34,401) we show that the time-on-task effect in digital reading is moderated by item difficulty and person skill. GLMMS revealed positive time-on-task effects for hard items and weak readers. For easy items the time-on-task effects were negative, and zero for skilled readers.

Investigating the Item-Position Effect in the PISA Assessment
Rianne Janssen, Qian Wu, Tiziana Lange, Dries Debeer, and Tiziana Lange, University of Leuven

Individual differences in the negative item-position effect have been found in PISA assessments. This position effect can be modeled as test-takers' persistence, or a change in examinee effort during testing. This study explores this interpretation, by relating the item-position effect to background variables and self-reported motivation.

Modeling Test-Effort for Low-Stakes Test: An IRTree Modeling Approach
Haiqin Chen, American Dental Association, Guangming Ling, Educational Testing Service, and Paul De Boeck, The Ohio State University.

This study proposes an extended IRTree model to capture different effort-related test taking behaviors defined by response time thresholds through outlier detection methods. We argue that this approach may help improve our understanding of test-taking behaviors associated with low-stakes standardized tests, as well as the assessment accuracy.

Wording Effects in TIMSS Motivational Scales Across Languages
Michalis Michaelides, University of Cyprus, Nicosia, Cyprus

The study examines the factorial structure of the TIMMS2011 student motivation scales in mathematics using survey data from five countries. A negative wording effect is present in models that fit the data adequately. For the 4th-grade study participants reading achievement scores appear to relate systematically to this wording effect.

Saturday April 18, 2015
4:05 PM - 6:05 PM, Valencia, Lobby Level, Paper Session, I10

Investigations in Examinee Guessing and Response Time

Session Chair: Elizabeth Stone, Educational Testing Service
Session Discussant: Kathleen Gialluca, Pearson VUE

Hierarchical Modeling of Item Responses and Response Times for Testlets
Suk Keun Im, University of Kansas, Lawrence, KS

The study will introduce new response time model to address testlet effects using real and simulated data. This study will try to quantify the estimation errors as a function of various testlet effects, in the context of different test conditions. The outcome will be compared with traditional response time model.

Response Model for Rapid-Guessing Behaviors in Data With Timing Information
Artur Pokropek, Polish Academy of Sciences (IFiS), Warszawa, Poland

In this presentation response specific mixture IRT model is proposed for rapid guessing behaviors. The model specification allows that some measurements for some subject might behave different than others but will follow model defined for one of distinct classes. Presented model is conceptually similar to the HYBRID model (Yamamoto, 1989).

Comparing Two Item Response Models That Incorporate Response Times
Heru Widiatmo and Daniel Wright, ACT, Inc., Iowa City, IA

Two measurement models, which use both responses and response times for calibrating abilities, are compared. The models are the hierarchical (van der Linden, 2006) and Q-diffusion (van der Maas et al., 2011) models. The ability estimates are compared with the known values and those estimated with the 3-PL IRT model.

A Test for Response-time Homogeneity of Item Responses
Paul De Boeck, The Ohio State University, Haiqin Chen, American Dental Association, and Minjeong Jeon, The Ohio State University

Local dependence of response time and accuracy implies that the difficulty of an item varies depending on the response time. It is a violation of response time homogeneity of item responses. A new Mantel-Haenszel test is proposed to detect such violations. The test is evaluated with a simulation study.

Detecting Errant Item Response Time Using Brownian Bridge Model
Fan Yang, Pearson/The University of Iowa and Stephen Dunbar, The University of Iowa

A new approach was proposed for detecting possible aberrant test behaviors based on item response time in high-stakes tests using Brownian Bridge Model. The preliminary simulation study results showed that this approach can be a reliable supplement to existing methods to detect and predict errant response time.

Semi-Parametric Item Response Functions in the Context of Guessing
Carl Falk and Li Cai, University of California, Los Angeles

We present a logistic function of a monotonic polynomial with a lower asymptote, allowing additional flexibility beyond the three-parameter logistic model. The item model is demonstrated on state math assessment data, and a strategy for choosing the order of the polynomial is demonstrated and tested.

Annual Meeting Program - Sunday, April 19, 2015

Sunday, April 19, 2015
5:45 AM–7:00 AM, Meet in the lobby of the InterContinental Hotel

NCME Fitness Run/Walk

Organizers:
Brian F. French, Washington State University
Jill van den Heuvel, Alpine Testing Solutions

Run a 5K or walk a 2.5K course in Chicago. Meet in the lobby of the InterContinental Hotel at 5:45AM. Pre-registration is required. Pickup your bib number and sign your liability waiver at the NCME Information Desk in the InterContinental Hotel, anytime prior to race day.

The event is made possible through the sponsorship of:

ACT

Alpine Testing

American Institutes for Research

American Institute of CPA

Applied Measurement Professionals, Inc.

Buros

College Board

CTB/McGrawHill

Educational Testing Service

GMAC

HumRRO

Law School Admission Council

National Board of Medical Examiners

National Center for the Improvement of Educational Assessment

National Council of State Boards of Nursing, Inc.

Pacific Metrics

SSATB

West Ed

Sunday, April 19, 2015
6:30 a.m.-7:30 a.m., Grand Ballroom Balcony, 8th floor

Yoga

Please join us for the first inaugural sunrise yoga sponsored by NCME. We will start promptly at 6:30 a.m. for one hour. Advance registration required ($10). NO EXPERIENCE NECESSARY. Just bring your body and your mind, and our instructor, Pierce (www.piercedoerr.com) will do the rest. Namaste.

Sunday April 19, 2015
8:15 AM - 10:15 AM, Empire Ballroom, 7th Floor, Paper Session, J1

Improving Proficiency Estimation

Session Chair: Rose Zheng, Pearson
Session Discussant: David Thissen, University of North Carolina Chapel Hill

Trait Estimation Effects on Proficiency Scores for End-of-Grade Examinations
Nurliyana Bukhari, Allison Ames, and Jonathan Rollins, University of North Carolina at Greensboro

The maximum likelihood (ML) estimator, frequently applied in latent trait parameter estimation by state departments of education, has limitations. The use of a weighted parameter estimator may provide substantial improvement over the ML procedure. This study aims to investigate the trait estimation effects on proficiency scores for the End-of-Grade exams.

The Performance of Robust Estimators in Ordinal Structural Regression Models
ChengHsien Li, UT Health Science Center at Houston, Houston

Robust maximum likelihood (MLR), robust unweighted least squares (ULSMV), and robust weighted least squares (WLSMV) have been considered to be superior to maximum likelihood when ordinal variables are analyzed. A Monte Carlo study is used to examine the effects of number of categories, level of asymmetric distributions, and sample size.

Improving Attribute Mastery Estimation in the LCDM With Covariates
Su-Pin Hung, National Cheng Kung University, Tainan, Taiwan

The present proposal aims to extend the LCDM using persons' covariates and to explore the effect of covariates on attribute mastery estimation within the framework of CDMs. Two simulation studies are designed to assess model parameter recovery by manipulating different numbers of attributes, Q-matrix structures, and sample sizes.

Toward an Optimal Proficiency Estimator
Peter Baldwin, National Board of Medical Examiners, Philadelphia, PA

Two kinds of errors may be said to characterize an estimator: random and systematic. This paper introduces an estimator of proficiency for the one-parameter item response theory model that minimizes these errors according to weighting functions specified by a practitioner. The proposed estimator performed well compared to other widely-used estimators.

Marginal Maximum Likelihood Estimation via the Discounted Likelihood Method
Charles Iaconangelo and Jimmy de la Torre, Rutgers, The State University of New Jersey

The discounted likelihood method (DLM) is proposed as an alternative to traditional MMLE using the EM algorithm. It employs a quasi-Newton algorithm to maximize the likelihood directly, avoiding the E-step and the computation of analytic derivatives required in Newton-Raphson. A simulation study illustrates the viability of the procedure.

Centering in MCMC Estimation of IRT Item Parameters
Leslie Hendrix, University of South Carolina, Columbia, SC

Centering for Markov chain Monte Carlo estimation has been used in other disciplines but it is not commonly used in IRT. This work shows that centering for the 3PL model significantly improves the comparability of estimates from MCMC and the Bayes Modal BILOG-MG procedure.

Sunday April 19, 2015
8:15 AM - 10:15 AM, Exchange, 11th Floor, Paper Session, J2

Performance Scoring Using Raters and Constructed Responses

Session Chair: Daniel Jurich, National Board of Medical Examiners
Session Discussant: Dan Bolt, University of Wisconsin Madison

A Comparison of Newly-Trained and Experienced Raters
Yigal Attali, Educational Testing Service, Princeton

Novice raters participated in a short (30-minute) training and certification program for evaluating essays. Performance of the newly-trained raters was compared to that of expert raters. Results showed scores from the two groups exhibited similar measurement properties. Implications for the importance of initial training and screening of raters are discussed.

Setting Meaningful Expectations for Clinical Learners and Preceptors Using Rubrics
Ulemu Luhanga, Laura McEwen, and Jane Griffiths, Queen's University-Kingston, Kingston, Canada

In clinical education, multi-source feedback (MSF) tools are used to collect information from multiple assessors. Although assessors may agree on performance, they tend to interpret MSF rating scales differently. To improve the utility of MSF tools, conversion of rating scales to rubrics may result in a shared frame of reference.

Handwritten vs. Typed: Effects on Essay Scores and Rater Cognition
Angelica Rankin, Stephen Dunbar, and Catherine Welch, University of Iowa, Iowa City, IA

Multiple methods were used to examine score differences between handwritten and word-processed essays. Six trained raters analytically scored 600 essays. MANOVA was used to examine score differences, and think-alouds and interviews were used to explore differences in rater cognition. Mode differences are highlighted and implications discussed.

Specific Agreement to Assess Rater Association
Anna Topczewski and Jennifer Beimers, Pearson, Ann Arbor, MI

In other fields, specific agreement has been used, in conjunction with Kappa, to assess rater associations. Specific agreement examines the agreement for a particular decision, such as rater agreement at a given score point. This study examines how specific agreement could be used to improved rater agreement analyses.

Analytic vs. Holistic Scoring of Identical Constructed-Response Items: Different Outcomes
Milja Curcin and Ezekiel Sweiry, Standards and Testing Agency, Department for Education, London, United Kingdom

A partial credit multi-facet Rasch model was used to answer the questions of whether points-based (analytic) and levels-based (holistic) scoring rubrics functioned interchangeably when scoring three-point constructed-response English reading comprehension items in the national tests for 11-year-olds in England, and whether the raters functioned interchangeably between these rubrics.

Sunday April 19, 2015
8:15 AM - 10:15 AM, Grand Ballroom, 7th Floor, Coordinated Session, J3

Innovative Perspectives on Common-Core Tests: PARCC & SBAC Compare Notes

Session Chair: S.E. Phillips, Consultant
Session Discussant: Mike Beck, BETA Inc.

Presenters: S.E. Phillips, Assessment Law Consultant, Laurie Wise, HumRRO, Derek Briggs, University of Colorado, Marty McCall, SBAC and Mike Beck, BETA

NCME members will be updated by consortia representatives about innovations in the tests being administered operationally in Spring 2015. A panel discussion will focus on 9 key areas including unique items and scoring, mode comparability, alignment, accommodations, security, performance standards, subscore reliability, fairness for diverse populations and costs.

Sunday April 19, 2015
8:15 AM - 10:15 AM, King Arthur, 3rd Floor, Paper Session, J4

Detecting Bias Across Special Populations

Session Chair: Xia Mao, Pearson
Session Discussant: Stephen G. Sireci, University of Massachusetts Amherst

Some Psychometric Consequences of Subpopulation Item Parameter Drift
Anne Corinne Huggins-Manley, University of Florida, Gainesville, FL.

This study hypothesizes that the presence of subpopulation item parameter drift is associated with bias in proficiency and scaling constant estimation, as well as subpopulation differences in dimensionality structure. It is demonstrated that these effects go beyond what can be understood from item parameter drift or differential item functioning analysis.

Who's On First? Gender Differences in Performance on SAT®-CR Items
Kay Chubbuck, W. Edward Curley, Teresa C. King, Educational Testing Service, Princeton, NJ

This paper summarizes quantitative and qualitative data concerning gender differences in performance on SAT®-CR material with sports and science content. Results indicate issues related to current methods of evaluating DIF statistics, the minimal impact of test-taker interest, and the possibility that passage length can mitigate gender differences in performance.

Gender DIF on Mixed Math Items Differing in Cognitive Demand
Ming-Chih Lan and Min Li, University of Washington, Seattle

Gender DIF items and patterns on math items differing in cognitive demand were examined by comparing OLR and poly-SIBTest methods. The study concluded that OLR and poly-SIBTest were consistent in identifying items as DIF and non-DIF but different in identifying the magnitude of DIF and types of DIF.

Comparing DIF Approaches to Investigate Home Language in an ELP Assessment
Shu-Jing Yen, Jennifer Renn, and Shauna Sweet, Center for Applied Linguistics, Washington, DC

This study compares exploratory and confirmatory DIF techniques to discern the potential benefit of creating item bundles based on expert linguistic analysis to determine whether performance on given items on high-stakes English language proficiency assessment varies across test takers from different language backgrounds after controlling for test takers' ability.

Identifying Bias Across Generated Mathematical Items of Varying Complexity
Clifford Hauenstein, Georgia Institute of Technology, Atlanta, GA

The current project addresses issues of assessment bias from the context of a test item generator. Specifically, uniform and non-uniform differential item functioning is assessed across items with various levels of cognitive-linguistic load. The target sample includes middle school English language learners; the reference sample includes native English speaking peers.

Sunday April 19, 2015
8:15 AM - 10:15 AM, Renaissance Ballroom, 5th Floor,
Coordinated Session, J5

Gathering and Evaluating Validity Evidence Based on Response Processes

Session Chair: Jose-Luis Padilla, University of Granada
Session Discussant: Bruno Zumbo, University of British Columbia

Evidence of response processes is one of five key sources of validity evidence (AERA/APA/NCME, 2014). This session will address how response process evidence varies in importance across testing situations and provide examples of how to gather and interpret this evidence for both validity evaluation and item development purposes.

Complementarity Between Cognitive Interviewing Findings and DIF Results: Enhancing Validation and Test Design

Jose-Luis Padilla and Isabel Benitez, University of Granada, Spain; Aura-Nidia Herrera and Jonathan-David Rico, National University, Colombia

DIF results are frequently hard to understand. Taking advantage of the complementarity between sources of validity evidence, this paper illustrates how findings from cognitive interviewing can be integrated with DIF results within a mixed-method research in order to build a validity argument, and inform test design

Using Response Process Evidence to Evaluate Language Demands in Academic Assessments

Ellen Forte, edCount, LLC

Developers of academic assessments must ensure that the language they use is relevant to the construct, accessible and comprehensible to examinees, and does not inappropriately influence students' ability to generate responses. This paper addresses these linguistic issues as they relate to score interpretation and use, validity evaluation, and item development.

Accounting for Affective States in Response Processes: Impact for Validation

Jacqueline P. Leighton, University of Alberta, Canada

The Standards (AERA, APA, NCME, 1999) are clear about the importance of response process data as a source of validity evidence for test and item score interpretation. The purpose of this paper is to summarize how affective/emotional processes influence test performance, then explain and propose reasons to include research on affective processes in validation arguments.

What Can Item Response Times Tell Us About DIF for English Learners?

Joshua Marland, Stephen G. Sireci, April Zenisky, and Duy Pham, University of Massachusetts Amherst

Response time differences and differences in response time engagement across ELs and non-ELs were used to interpret DIF. The results provide insight into how differential motivation and persistence affect students' performance on items and illustrate how item response times provide validity evidence based on response processes to help interpret DIF.

Sunday April 19, 2015
8:15 AM - 10:15 AM, Seville Ballroom East, Lobby Level, Invited Session, J6

Exploring the Implications of the "Fairness" Chapter of the 2014 Standards for Educational and Psychological Testing

Session Chair: Meagan Karvonen, CETE, University of Kansas
Session Discussants: Edynn Sato, Pearson, Peggy Carr, NCES, and Brian Gong, Center for Assessment

The revised AERA/APA/NCME Standards for Educational and Psychological Testing were published in summer 2014. A dramatic shift in the Standards is the addition of a foundational chapter on "Fairness" and the removal of chapters on language diversity and students with disabilities. In this session, presenters deeply involved in the development of the Standards, and specifically in the chapter on fairness, will share their reflections on the development process and decisions made. Discussants from the perspectives of a state assessment contractor, NAEP representative, and organization that works with states on technical adequacy of assessments provide their reflections.

Presenters:

Overview and Introduction to the Revised Standards
Barbara Plake, Co-Chair, Joint Committee for the Revised Standards

Fairness Chapter Changes and Implications
Linda Cook, Member, Joint Committee for the Revised Standards

Sunday April 19, 2015
8:15 AM - 10:15 AM, Seville Ballroom West, Lobby Level, Paper Session, J7

Multidimensional Item Response Theory

Session Chair: Alejandra Garcia, University of Massachusetts
Session Discussant: Rich Patz, ACT

Assessing Growth Using Multidimensional Item Response Theory
Zhen Wang, Educational Testing Service and Lihua Yao, Defense Manpower Data Center

The major goal of this study is to illustrate the use of several different psychometric models (MIRT vs. non-MIRT) to link the tests across years. Using a college level learning outcome assessment, we will demonstrate how different models can impact students' growth calculation at both overall domain and subdomain level.

MIRT Analysis of Longitudinal Assessment When Constructs Vary Over Time
Hi Shin Shim and James Roberts, Georgia Institute of Technology, Atlanta, GA

This paper illustrates a multidimensional item response theory model, called the Sprout Model (SM), for measuring individual change in longitudinal assessments when latent traits are not necessarily invariant across time. SM parameter recovery is described, and the utility of the model is demonstrated with an analysis of data from ECLS-K.

An IRT Model for Multidimensional Ranking Data in Ipsative Tests
Xue-Lan Qiu, Wen-Chung Wang, Hong Kong Institute of Education and Shungwon Ro, IBM Software Group

By design, multidimensional raking data are ipsative and cannot measure the absolute level of latent traits. In this study, we developed an IRT model for multidimensional ranking data where different statements measure different latent traits. We conducted a brief simulation study to evaluate parameter recovery and provided an empirical example.

Modeling Growth With Adaptive Longitudinal Large-Scale Assessments
Jiahe Qian, Educational Testing Service, Princeton, NJ

Growth models for longitudinal data can benefit from use of demographic information, and to this end, the 2PL MIRT model is employed to build latent logistic regression models for the National Education Longitudinal Study of 1988 (NELS:88). The inquiry aims to improve modeling change of person related parameters.

A New Mixture MIRT Approach for Measuring Response Styles
Lale Khorramdel-Ameri and Matthias von Davier, Educational Testing Service, Princeton

A mixture MIRT approach to measure and correct for response styles (RS) is presented. It is shown that RS can be measured unidimensionally and differentiated from trait-related responses, and that different respondent groups show different kinds of RS. Results are validated using correlations between noncognitive scales and cognitive test scores.

Assessing Model-Data Fit for Compensatory and Non-Compensatory MIRT Models
Leanne Freeman and Bo Zhang, University of Wisconsin, Milwaukee

The objective is to provide evidence of statistical options for model-data fit comparisons between compensatory and non-compensatory multidimensional item response theory (MIRT) models. The Clarke and Vuong statistics will be studied in various conditions through Monte-Carlo simulation. Effectiveness will be evaluated by both Type I error and statistical power.

Sunday April 19, 2015
8:15 AM - 10:15 AM, Toledo, 5th Floor, Coordinated Session, J8

Delivering the National Assessment on Tablet: Psychometric Challenges and Opportunities

Session Chair: Andreas Oranje, Educational Testing Service
Session Discussant: Bill Tirre, NCES

In order to maintain a valid, reliable, and fair indicator of what students know and can do, the National Assessment of Educational Progress (NAEP) is in the process of transitioning from paper to technology based assessment, in particular tablet administration. This transition entails a number of interesting psychometric questions that need to be answered:

1. Can trends be maintained across two modes: paper and tablet? If so, under what conditions and with what kind of design in place?
2. Where and among what student groups are digital literacy gaps and how do those play a role in this transition?
3. Technology provides the examinee with new ways to interact with assessment materials. How do these affordances affect maintaining trend in the short and long terms?
4. How can the affordances of technology be used to improve administration designs and, subsequently, improve the assessment?
5. What kind of statistical analysis models should be utilized, particularly in a group score context, and how should decisions about the fidelity of the trend be made? How do additional data sources (e.g., click streams, log files, time stamps) play a role in that decision?

Introduction to Delivering the National Assessment on Tablet: Psychometric Challenges and Opportunities
Janeen McCullough, Educational Testing Service

The first paper sets the stage and will provide an introduction of the general transition design as well as some subject (Reading, Math, and Science) specific aspects while introducing terminology and definitions that subsequent papers will build upon.

Digital Literacy and Performance Gaps in Computer-based Assessments
Ting Zhang, Young Yee Kim, George Bohrnstedt, Markus Broer, American Institutes for Research; Qingshu Xie, MacroSys

A key condition for the transition will be addressed: digital literacy and performance gaps, presenting the results of various recent mode studies including NAEP grade 8 computer-based assessments (CBAs) in writing and mathematics to investigate the relationship between student CBA performance and their digital literacy, especially their computer familiarity.

Developing and Trans-Adapting Items for Technology-Based Assessment
Rebecca Moran, Hilary Persky, and Gloria Dion, Educational Testing Service

Key aspects of item and task development and trans-adaptation related to the psychometric claims the program intends to make will be covered, striking a balance between maintaining trends and introducing technology-based tools and interactive item and task types with the goal to improve measurement.

Distributing Tasks and Items in Technology Based Group Score Assessments
Longjuan Liang and Ed Kulick, Educational Testing Service

We will focus on administration design aspects of technology based assessments, including the use of adaptive designs and random booklet assignment, and how those affect the ability to maintain trends, drawing heavily on simulation work.

Analysis Methods for TBA Group Score Assessments
Zhan Shu and Katherine Castellano, Educational Testing Service

We will address statistical modeling questions associated with a transition both in terms of the types of models used and the kind of inferences that can be made based on the design aspects presented in the previous papers. Results from recent technology based assessments will be used to illustrate the principles that were applied.

Sunday April 19, 2015
8:15 AM - 10:35 AM, Valencia, Lobby Level, Invited Session, J9

Awards Session
Session Chair: Lei Wan, Pearson
Session Discussant: Laine Bradshaw, University of Georgia

Jason Millman Award
The Innovative Applications of Response Time in Detecting Aberrant Behaviors in Standardized Testing
Chun Wang, University of Minnesota

Two mixture hierarchical models based on response accuracy and response times will be introduced, and demonstrate how the new models can be used to detect aberrant behaviors--rapid guessing and cheating behaviors. The performance of the new model based approach is also compared to residual-based fit indices.

Alicia Cascallar Award
Covariate and Mixture Extensions of Diagnostic Classification Models
Yoon Soo Park, University of Illinois at Chicago

Diagnostic classification models (DCMs) classify examinees into attribute mastery profiles. This study presents extensions of DCMs that incorporate mixture distributions to examine differential attribute functioning among latent subgroups. Covariates are specified at the attribute and higher-order latent trait levels to explain differences in attribute structures, response probability, and latent classification.

Brenda Loyd Award
Estimation of Complex Generalized Linear Mixed Models for Measurement and Growth
MinJeong Jeon, The Ohio State University

In this talk, I will present my dissertation that addresses estimation methods and applications of complex generalized linear mixed models for measurement and growth. I will briefly explain two noble maximum likelihood techniques that I developed – variational maximization-maximization (VMM) and Monte Carlo local likelihood (MCLL) algorithms. I will also present a proposed autoregressive growth model and its implication in measurement.

Bradley Hansen Award
A Multilevel Testlet Model for Mixed-Format Tests
Hong Jiao, University of Maryland, College Park

This research project proposes a multilevel testlet model for mixed-format tests consisting of both dichotomous and polytomous items. This modeling approach can tackle multiple psychometric issues such as dual local dependence due to item and person and complex sampling in testlet-based in applying conventional item response theory (IRT) models. Parameter estimation accuracy will be evaluated under simulated study conditions. Further the proposed model will be compared with three competing models in terms of parameter recovery for mixed-format tests.

Sunday April 19, 2015
10:35 AM - 12:05 PM, Empire Ballroom, 7th Floor, Coordinated Session, K1

Multiple Facets of an Assessment With Collaborative Problem Solving Tasks

Session Chair: Jiangang Hao, Educational Testing Service
Session Discussant: Mengxiao Zhu, Educational Testing Service and Yigal Rosen, Pearson

Presenters: Alina von Davier, Lei Liu, Patrick Kyllonen, Saad Khan, and Jiangang Hao, Educational Testing Service

In this symposium we present the first set of results of various research strands conducted on understanding the implications of including collaborative problem solving tasks (CPS) in simulation based assessments.

Sunday April 19, 2015
10:35 AM - 12:05 PM, Exchange, 11th Floor, Coordinated Session, K2

Designing Next-Generation Assessments of Student Learning Outcomes in Higher Education

Session Chair: Katrina Crotts Roohr, Educational Testing Service
Session Discussant: Javarro Russell, Educational Testing Servic

This session discusses assessments for five college-level learning outcomes: critical thinking, written communication, oral communication, quantitative literacy, and intercultural competence. We offer a comprehensive review of each competency and discuss important assessment considerations when designing next-generation assessments. Measuring these competencies has important implications for higher education institutions and the workforce.

Assessing Critical Thinking in Higher Education
Ou Lydia Liu, Lois Frankel, and Katrina Crotts Roohr, Educational Testing Service

The importance of critical thinking skills is widely recognized by higher education institutions and employers in a global economy. This paper provides a comprehensive review of current definitions and assessments of critical thinking and discusses challenges and assessment considerations for designing a next-generation critical thinking assessment for college students.

Assessing Written Communication in Higher Education: Review and Recommendations
Jesse R. Sparks, Wyman Brantley, Yi Song, and Ou Lydia Liu, Educational Testing Service

The ability to communicate effectively in writing is widely acknowledged as a critical competency for academic and workforce success. This paper reviews existing definitions and assessments of writing skills, and proposes a research-based construct definition and framework for designing next-generation assessments of written communication as a higher education outcome.

Oral Communication in Higher Education: Existing Research and Future Directions
Katrina Crotts Roohr, Liyang Mao, Vinetha Belur, and Ou Lydia Liu, Educational Testing Service

Oral communication has been identified as an important skill for college graduates. By synthesizing the existing oral communication frameworks and assessments, this paper provides an operational definition for a next-generation oral communication assessment in higher education. Challenges of designing such an assessment are also discussed.

Assessing Quantitative Literacy in Higher Education
Katrina Crotts Roohr, Edith Aurora Graf, and Ou Lydia Liu, Educational Testing Service

Quantitative literacy is the ability to interpret and communicate numbers and mathematical information throughout everyday life, and is recognized as an important skill in higher education and the workforce. This paper synthesized existing frameworks and assessments and proposed an operational definition for a next-generation assessment, discussing assessment considerations and challenges.

Intercultural Competence in Higher Education: Current State and Future Directions
Meagan Caridad Arrastia, Florida State University; Joseph A. Rios, Ou Lydia Liu, Liyang Mao, Lauren Carney, and Meghan W. Brenneman, Educational Testing Service

Intercultural competence (ICC), the ability to communicate across cultures, has become a valued learning outcome for current college students. By reviewing the existing ICC frameworks, this paper discussed the considerations of assessing ICC skills and developed an operational definition for a next-generation ICC assessment within a higher education context.

Sunday April 19, 2015
10:35 AM - 12:05 PM, King Arthur, 3rd Floor, Coordinated Session, K3

Constructing a Vertical Scale Under Linked Scaling Tests Design

Session Chair: Wei Tao, ACT, Inc.
Session Discussant: Deb Harris, ACT, Inc.

An Overview of the Linking Designs and Statistical Methods
Wei Tao and Andrew Mroch, ACT, Inc.

To inform the development of a vertical scale, several simulation studies were conducted. Three linking designs based on random equivalent groups, common items, and external anchor items are introduced. Several IRT and non-IRT approaches to data analyses are presented.

Linking Individual Scaling Tests Using Non-IRT Methods
Dongmei Li, ACT, Inc.

Two non-IRT methods of vertical scaling were explored— the Thurstone method and an ad hoc method involving predictions of the whole scaling test raw scores by equipercentile or linear linking across the individual scaling tests. Results were compared across these methods and variations within each method.

Linking Individual Scaling Tests Using IRT Methods
Troy Chen and Wei Tao, ACT, Inc.

To place the item parameter estimates on the same scale after calibrating separate scaling tests, the scale transformation and fixed item parameter estimation approaches were considered. Evaluation criteria such as growth patterns and correlations between parameters and estimates were used to assess these methods.

Different Designs and Methods for Scaling
Yu Fang, ACT, Inc.

This study investigated various approaches to placing on-grade test scores on a vertical scale under the scaling test design presented in this symposium. Assumptions underlying different analysis methods and characteristics of a vertical scale such as constant conditional standard error of measurement (CSEM) will be discussed.

Sunday April 19, 2015
10:35 AM - 12:05 PM, Renaissance Ballroom, 5th Floor,
Coordinated Session, K4

Do Interruptions During Online Testing Impact the Examinee Scores?

Session Chair: Craig Mills, CTB McGraw Hill
Session Discussants: Walter (Denny) Way, Pearson and Neal Kingston, University of Kansas

The number of interruptions during online testing is on the rise. This session focuses on the determination of impact of interruptions on the examinees' performance. Impact can be determined at an overall level, school level or individual level. Several approaches are proposed for determining the impact at the different levels.

Practical Considerations When Interruptions Occur
J.P. Kim, ACT, Inc.

This presentation will discuss several practical considerations when interruptions occur. These considerations include options for resuming the test, collecting available resources for conducting follow-up analyses, and determining the unit of analysis and statistical methods. Considerations will be discussed with examples from an operational test.

Overall Assessment of the Impact of Interruptions
Dong-In Kim, McGraw-Hill Education CTB; Sandip Sinharay, Pacific Metrics; Ping Wan, Seung W. Choi, and Litong Zhang, McGraw-Hill Education CTB

This presentation will focus on several approaches for performing an overall assessment of the impact of interruptions. The approaches involve statistical hypothesis testing, statistical methods for matching, and IRT. The methods will be applied to data from the 2013 Indiana state tests.

School-Level Assessment of the Impact of Interruptions
Arthur A. Thacker and Bethany H. Bynum, Human Resources Research Organization

This presentation will discuss several approaches for performing a school-level assessment of the impact of interruptions and the constraints associated with performing such an assessment. Three different methodological approaches will be discussed using data from 2013 Minnesota Comprehensive Assessment (MCA).

Individual-Level Assessment of the Impact of Interruptions
Sandip Sinharay, Pacific Metrics; Ping Wan, Seung W. Choi, and Dong-In Kim, McGraw-Hill Education CTB

This presentation will focus on several approaches for performing an individual-level assessment of the impact of interruptions. The methods involve ideas from statistical hypothesis testing, linear regression, and IRT. The methods will be applied to data from the 2013 Indiana state tests.

Sunday April 19, 2015
10:35 AM - 12:05 PM, Seville Ballroom East, Lobby Level,
Coordinated Session, K5

Current Issues in Test Assembly

Session Chair: Jonas Bertling, Educational Testing Service
Session Discussant: Wim van der Linden, CTB

This symposium addresses the assembly of test forms with defined item difficulties that are not only reliable and valid but also reasonably short. We will present state-of-the art applications of rule-based AIG and solutions to important issues concerning test assembly that might easily be overlooked when focusing on item generation.

Improving Cognitive Tests by Rule-Based Generation of Test Items and Distracters
Jonas P. Bertling, Maria Bertling, and Jim Fife, Educational Testing Service, Princeton, NJ

This paper presents an application of AIG where both items and distracters were generated based on rule-based principles. Large-sample analyses showed item difficulties were predicted from the underlying AIG model and that a systematic distracter generation and analysis improved test validity.

Optimizing the Assembly of Parallel Test Forms via Linear Programming
Jonathan Weeks, Educational Testing Service, Princeton, NJ

This empirical study presents an application of linear programming to the assembly of parallel forms of fluid reasoning measures. The assembly considers parallelism with respect to test information, test characteristic curves, and content representation. Additional constraints are added regarding item and total response time.

Using Response Time Data to Reduce Testing Time in Cognitive Tests
Maria Bertling and Jonathan Weeks, Educational Testing Service, Princeton, NJ

The authors argue that shorter, yet reliable and valid, test forms can be designed by incorporating response-time data in ability estimates. The incremental contribution to the precision of ability estimates and to test information is discussed in a large sample of a highly educated population.

Improving Reasoning Tests by Use of Instant Feedback
Achim Preuss, cut-e Group, Germany and Katharina Lochner, cut-e Consulting Singapore, Singapore

Performance on ability tests is influenced by participants' perception of their own performance. We showed that feedback after the example section does not impact performance on an online reasoning test, whereas instant feedback during test completion impacts performance and processing style and can improve test validity.

Sunday April 19, 2015
10:35 AM - 12:05 PM, Seville Ballroom West, Lobby Level,
Coordinated Session, K6

Toward More Robust Automated Essay Scoring Models

Session Chair: Chaitanya Rameneni, Educational Testing Service
Session Discussant: Sue Lottridge, Pacific Metrics

Automated scoring model performance can be sensitive to the quality of responses used for training and evaluating the model. This session presents studies that investigate new methods for effectively detecting and filtering aberrant responses to yield more robust scoring models.

Optimal Design to Improve Essay Selection and Scoring Generalizability
Nicholas Dronen and Peter W. Foltz, Pearson

Optimal design of experiments provides a novel approach to finding essays that cover the spectrum of features related to the automated scoring models. We describe results of studies that explore the generalizability of the method and its applicability to selecting essays.

The Short, Irrelevant and Odd; Automated Detection of Aberrant Essays
Anat Ben-Simon, Yael Safran and Yoav Cohen, National Institute for Testing and Evaluation, Jerusalem, Israel

The study examines the efficiency of an algorithm developed for the detection of aberrant essays written in the Hebrew and Arabic languages. The algorithm uses 10 criteria to classify essays as proper or gibberish. The classification accuracy obtained, was 99.6% and 96.7% for essays in Hebrew and Arabic respectively.

Automatic Detection of Nonscorable Essays
Yinghao Sun, The Ohio State University and Vincent Kieftenbeld, McGraw-Hill Education CTB

Student responses that deviate substantially from the requirements of a constructed-response item often cannot be scored according to the rubric. Automated scoring systems need to successfully detect such nonscorable responses. We investigated different features, classifiers, and ensemble learning methods to detect nonscorable essays in a recent large-scale assessment.

Using Automated Features to Detect Aberrant Prompts and Responses
Frank Williams and Chaitanya Ramineni, Educational Testing Service

New prompts written for operational use with potential to elicit responses inappropriate for automated scoring can be a threat to reliability and validity of automated scores, and result in a degradation of performance of an operational scoring model. This study explores the use of statistical inferential procedures on automated features to detect such prompts prior to release for use with the operational scoring model.

Sunday April 19, 2015
10:35 AM - 12:05 PM, Toledo, 5th Floor, Coordinated Session, K7

Detection and Solutions of Aberrant Performances of Automated Scoring Systems

Session Chair: Christy Schneider, CTB McGraw-Hill
Session Discussant: Christy Schneider, CTB McGraw-Hill

We will discuss a wide range of problems which cause aberrant performance in automated essay scoring and spoken response scoring (e.g., technical difficulties, characteristics of individual examinees, responses, items, and item-types). We will provide methods to identify and correct the aberrant performance while improving the validity of automated scores.

What Makes the Automated Speech Scoring System Off-Target?
Guangming Ling and Su-Youn Yoon, Educational Testing Service

We filtered a subset of speaking responses with higher probability of having a greater machine scoring errors (HMSD) by considering characteristics of speakers, tasks, and machine-generated speech features, and had them scored by human raters; we found that this approach actually improved the scoring quality of the machine scores.

Development of a Non-Scorable Test Detection System for English Language Learner Assessment
Xin Chen, Angeliki Metallinou, Yuan Zhao-D'Antilio, and Jian Cheng, Pearson

This paper describes a method to develop a system to detect non-scorable tests, i.e., tests that cannot be confidently scored automatically, in automated spoken language proficiency assessment. The detection system was developed based on a dataset from a large scale, high-stakes English language proficiency test, administered to K-12 ELL students in a U.S. state. Implications of a potential application of such system for different stakeholders will be discussed.

Using Automated Generic Scoring Models to Identify Deviant Prompts
John Mattar, AICPA, Chaitanya Ramineni, Educational Testing Service, Aster Tessema, AICPA, Chen Li, Educational Testing Service, and Matthew Schultz, AICPA

Automated generic scoring models are used to identify prompts for which the scoring model does not meet desired performance standards, and implications for prompt development and comparability are discussed.

Automated Scoring of Source-Based Writing With Rare (Scientific) Vocabulary
F. Jay Breyer, John Blackmore, and Laura Ridolfi-McCulla, Educational Testing Service

Automated scoring of writing tasks with technical/scientific terms in sources can show separation between human and machine scores, where humans give higher grades compared to machines. A method is described and evaluated that identifies and reduces score separation for writing tasks with sources using rare vocabulary in a generic model.

Sunday April 19, 2015
10:35 AM - 12:05 PM, Valencia, Lobby Level, Paper Session, K8

Using Validity Evidence in Diverse Settings

Session Chair: Jenna Copella, Pearson
Session Discussant: Michael Kane, Educational Testing Service

Development of a High-Stakes Exam for Architecture Graduates in Mexico
Laura Delgado-Maldonado, Jorge Hernández-Uralde, Rafael Sánchez-Mayorga, Instituto Nacional para la Evaluación de la Educación; Melchor Sánchez-Mendiola, UNAM Faculty of Medicine; and Eduardo Ramírez-Díaz, CENEVAL

This paper describes the development and implementation of a high-stakes exam for Architecture graduates in Mexico, using the design of an architectonic project as strategy for assessment of complex performance. The test development process, accumulation of validity evidence, judges' training, instrument development and initial application data are described.

Longitudinal Confirmatory Factor Analysis for the SGL Student Evaluation Form
Cigdem Alagoz-Ekici, Scott Richardson, V. Thomas Gaddy, Lynn Doster, Gerald Crites, Brett Szymik, and Eve Gallman, University of Georgia, Athens, GA

Meaningful measurement of medical student competencies requires having a valid instrument that measures these competencies/behaviors over time. The validity of a Small Group Learning Student Evaluation instrument was evaluated by testing the measurement and structural invariance using longitudinal confirmatory factor analysis with longitudinal data from first- and second-year medical students.

A Validation Study of Personality Assessment Inventory (PAI) for Korean Delinquent Juvenile Probationers
Hye-Sook Park, Honam University, Gwangju, Republic of Korea

This study validates Personality Assessment Inventory for Adolescent (PAI_A) scale using both classical test theory and Rasch measurement model. Data were collected from Korean juvenile delinquents under the supervision of Korean criminal justice systems in Korea.

The Validation of the Mathematics and Science Instructional Logs
Elizabeth Greive, Carrie Lee, and Temple Walkowiak, North Carolina State University, Raleigh, NC

Instructional logs provide an innovative means for measuring instructional practices. This proposal describes the evidences of validity collected for the MSI Logs, which are designed for teachers to report their daily practices. Inter-rater agreement statistics and cognitive interview findings are described and suggestions are made to improve the measures.

Sunday April 19, 2015
12:25 PM - 1:25 PM, Camelot, 3rd Floor, Electronic Board Session, Paper Session, L1

Electronic Board #1-4
Coordinated Session

Missing Data in Large-Scale Assessments
Tanya Longabach, Excelsior College; Shenghai Dai, Xin Yuan, and Yan Zhou, Indiana University

This group of studies explores different methods of missing data handling in large-scale assessments. Causes of different types of missingness, including omitted and not reached items, are examined, and parameter estimation bias is compared between scoring items as incorrect, ignoring missing items, scoring as partially correct, and various imputation methods.

Non-Response Models for Test Speededness
Yu-Wei Chang, Department of Statistics, University of Illinois at Urbana-Champaign; Nan-Jung Hsu and Rung-Ching Tsai, National Taiwan Normal University

We aim to investigate suitable models, with non-response as a response category, for speeded data. A popular non-response model is suggested for specific speededness data under a strong assumption. To relax the assumption, we extend the non-response model and work out the estimation via the penalized quasi-likelihood approach.

A New Modeling of Expert-Defined Multiple-Level Attributes With DINA Model
Wei Tian, National Assessment Center for Education Quality, MOE, China, Tao Xin, Beijing Normal University, Beijing, China, and Jiahui Zhang, Michigan State University

Multiple-level attributes are commonly defined to measure the realization of standard-based educational objectives. Such polytomous attributes as part of Q matrix will be more directly modeled into DINA. In addition, a statistical computing method was proposed to make it practicable. Its practical usefulness was examined with simulated and real-data.

Rasch Model Parameter Recovery With a Conditional Pseudo-Likelihood
John Willse, Jonathan Rollins and Saed Qunbar, University of North Carolina Greensboro

This study examines parameter recovery from polytomous Rasch models when using variants of conditional pairwise maximum pseudo-likelihood first described by Andrich and Luo (2003). This pseudo-likelihood approach has features that may make it preferable to other techniques when sample sizes are small or the number of thresholds is large.

Electronic Board #5
A Restricted Bi-Factor Model
Yu-Feng Chang and Mark Davison, University of Minnesota, Minneapolis, MN

This study proposes a restricted bi-factor model, which provides a way to identify examinees' weaknesses and strengths using the subscores corresponding to specific factors. A simulation study was designed to evaluate how accurately this model identifies the examinees' significant strengths or weaknesses.

Electronic Board #6
An Exploratory Study of Speededness Effects on IRT Model Fit
Lu Wang, ACT, Min Wang, The University of Iowa, and Troy Chen, ACT

This study investigates four approaches to identifying speededness and their impacts on IRT model-data fit under various IRT models. Real data from a large-scale test are employed. The results of this study will provide guidelines for practitioners when speededness effects are a concern.

Electronic Board #7
An Explanatory Longitudinal Multilevel IRT Approach to Instructional Sensitivity
Alexander Naumann, Jan Hochweber, and Johannes Hartig, German Institute for International Educational Research (DIPF)

We propose an explanatory longitudinal multilevel IRT model to evaluate items' instructional sensitivity. The model allows quantifying sensitivity to the instructional context and relating it to instructional measures. Results suggest that the model works well in its application to empirical data. Sensitivity was found to be related to instructional measures.

Electronic Board #8
Making National Assessments Diagnostically Useful: A Primary Education Mathematics Example
Daniel Van Nijlen and Rianne Janssen, KU Leuven, Leuven, Belgium

A test on the mastery of the four basic operations of arithmetic was analyzed using the Bayesian inference for binomial proportion model, a conjunctive cognitive diagnostic model with low computational demand. It is shown that the analyses are a valuable addition to the standard way of reporting on national assessments.

Electronic Board #9
Sequential Analysis for Detecting Learning in Cognitive Diagnosis
Sam Ye, Georgios Fellouris, Jeff Douglas, and Steven Culpepper, University of Illinois - Urbana Champaign, Champaign

Cognitive diagnosis models have been developed for assessing mastery and non-mastery of a vector of skills or attributes. In an e-learning environment, or in other settings, learning is a primary objective, and diagnostic models should detect this. We introduce change-point detection techniques from sequential analysis for this purpose.

Electronic Board #10
Comparing Longitudinal and Cross-Sectional School Effect Estimates in Postsecondary Education
Doris Zahner, CAE and Jeffrey Steedle, Pearson

Some universities administer standardized tests to estimate school effects on student learning. Many of them choose to gather cross-sectional rather than longitudinal data. This study compares longitudinal and cross-sectional school effect estimates using data from the CLA. Seven different statistical models were applied to the data.

Electronic Board #11
Are Fit Indices Biased in Favor of Bi-Factor Models?
Kari Hodge and Kevin Wells, Baylor University, Temple, TX

We tested the hypothesis that there is statistical fit bias favoring the bi-factor model by comparing correlated factors, higher-order, and bi-factor models via Monte Carlo simulations. The bi-factor model fit better even when the true model was not a bi-factor model. Each model fit the data well in absolute terms.

Electronic Board #12
Detecting Person Fit for Cognitive Diagnostic Assessment
Ying Cui, University of Alberta, Edmonton, Canada

This study examined ways of detecting person misfit for cognitive diagnostic assessments (CDA). We investigated whether lz, developed under the IRT, can be extended for use in CDAs. We also introduced a new statistic. Simulated and real data were used to compare the power of lz and our new statistic.

Electronic Board #13
Evaluating Dimensionality Assessment Procedures in Complex-Structure Noncompensatory Framework
Xiaying Zheng, Hong Jaio, and Qiwen Zheng, EDMS, University of Maryland, College Park, College Park, MD

Most dimensionality assessment procedures assume a simple-structure compensatory MIRT framework. This assumption does not always hold in real texting situations. This study examines the performance of seven common dimensionality assessment procedures with complex-structure noncompensatory MIRT models, and evaluates the applicability of these procedures in such context.

Electronic Board #14
Assessing Two Methods of Q-Matrix Validation for the DINA Model
Hueying Tzou, National University of Tainan, Jimmy de la Torre, Rutgers, The State University of New Jersey, Yi-Fang Wu, University of Iowa, and Ragip Terzi, Rutgers, The State University of New Jersey

When implementing CDMs, we often concern about model-data fit, especially the appropriateness of the Q-matrix. The current study focuses on two empirically based methods, the delta-method and the gamma-method, and examines the effects of the Q-matrix modification on the improvement of model-data fit and the correct classification of examinees.

Electronic Board #15
Diagnostic Test Designs: Multiple One-Attribute Models Versus One Multiple-Attribute Model
Laine Bradshaw, University of Georgia, Athens, GA

For formative assessment systems, diagnostic classification models (DCMs) can efficiently identify areas where students struggle. However, DCM estimation complexity increases quickly as the number of attributes measured (i.e., dimensionality) increases. We investigate using separate, single-attribute DCMs in lieu of a single, multiple-attribute DCM when data is insufficient for high-dimensional estimation.

Electronic Board #16
A Comparison of Clinical Classification Accuracy Using CDM and MIRT
Xuechun Zhou, Pearson, Liyang Mao, Michigan State University, Ou Zhang, Pearson, and Xin Luo, Michigan State University

This study examined clinical classification accuracy in an empirical setting with real data using the CDM and M2PL model. The results indicated that final model selection is determined by the purpose of clinical screening. In addition, the combination of the test scaled scores improved the correct classification rates significantly.

Electronic Board #17

Comparison of Marginal Item Response Theory Model Item-Fit Indices

Adrienne Sgammato and John Donoghue, Educational Testing Service, Princeton, NJ

Performance of several item-level IRT model fit indices based on residuals of one-way to four-way margins was evaluated in the presence of missing data. Additionally, item-level fit measures based on observed data were calculated. Type I error and power of these measures were compared to those based on model residuals.

Electronic Board #18

The Robust Sandwich Variance Estimators for the DINA Model

Jung Yeon Park, Matthew Johnson, and Young-Sun Lee, Columbia University, New York

Jackknife resampling is considered one of the most straightforward techniques to analyze large-scale data that use complex sampling designs (e.g. TIMSS). However, jackknife resampling has also been noted to be computationally inefficient. This study aims to develop sandwich estimators to generate more robust and efficient variance estimators in DINA model.

Electronic Board #19

Linear and Nonlinear Modeling of Item Position Effects

Chansuk Kang and Anthony Albano, University of Nebraska, Lincoln, NE

This study examines how different representations of item position can impact how the relationship between item position and item difficulty is expressed. Different explanatory IRT models are formulated with position as an item covariate, and these models are demonstrated using PISA 2009 reading data for the U.S.

Electronic Board #20

Multilevel Graded Response Testlet Model With Complex Sampling Designs

Dandan Liao and Hong Jiao, University of Maryland, College Park, MD

In practical settings, the assumptions of standard item response theory models might be violated. This study intends to generalize dichotomous multilevel testlet models to polytomous multilevel testlet models to account for dual dependence in testlet-based assessments with clustered samples. A multilevel graded response model (GRM) is evaluated in simulated conditions.

Electronic Board 21

Examining IRT Modification Index Test of Common IRFs Across Groups

John Donoghue, Educational Testing Service and Catherine McClellan, Clowder Consulting LLC

Evaluating the assumption of common IRF is an important issue in real-world applications of IRT. Procedures currently in use make questionable assumptions, and weak support from software exacerbates the problem. This paper examines the Lagrange Multiplier test to evaluate the assumption of invariant IRF in two populations.

Electronic Board #22

MIRT Classification Accuracy: Effects of Ignoring the Partially Compensatory Nature

Janine Buchholz, DIPF, and Joseph Rios, Educational Testing Service, and Johannes Hartig, DIPF, German Institute for International Educational Research

The model most commonly employed to estimate within-item multidimensionality is compensatory. However, numerous examples in educational testing suggest partially compensatory relations among dimensions. This simulation study aimed at evaluating the impact on theta estimates when the compensatory model is applied incorrectly. Findings imply only negligible effects on classification accuracy.

Sunday April 19, 2015
12:25 PM - 1:55 PM, Empire Ballroom, 7th Floor, Paper Session, L2

Subscore Reporting

Session Chair: Michelle Croft, ACT, Inc.
Session Discussant: David Shin, Pearson

Effects of Subgroup Ability Distributions on Subscore Reporting
Changjian Wang, Pearson, Lingyun Gao, ACT, and Lei Wan, Pearson

This proposed simulation study compares and evaluates the feasibility and effectiveness of three procedures for subscore reporting. The factors under consideration include sample ability distributions, performance groups used for augmentation/calibration, and correlation among subtests. Reliability, MSE, and PRMSE will be used in the evaluation of results.

Enhanced Subscale Reporting Using Unidimensional Item Response Theory
Robert Keller, Measured Progress, Lisa Keller, University of Massachusetts, and Michael Nering, Measured Progress

A method of improving subscale score reliability using unidimensional IRT models is investigated. Utilizing independent scales for each subscale, a method for scale creation is investigated through simulation study. A real data case study is also performed.

Impact of Score Report Design on Subscore Interpretation and Use
Sarah Carroll and Andrew Dwyer, Castle Worldwide, Inc., Morrisville, NC

This study examines how score report design elements affect candidate interpretation and use of subscores, and whether that relationship is affected by the level of agreement between candidate self-perceptions of subdomain knowledge and actual subdomain test performance.

Sunday April 19, 2015
12:25 PM - 1:55 PM, Exchange, 11th Floor, Paper Session, L3

Innovations in Teacher Evaluation

Session Chair: Heather Buzick, Educational Testing Service
Session Discussant: Erika Hall, Center for Assessment

Stakes Matter: Using Standardized Student Assessments for Teacher Evaluation
David Rutkowski and Justin Wild, Indiana University, Bloomington, IN

Employing an experimental study design in this study, we find that when 8th grade students were told that either their grades or their teacher's employment status were at stake, achievement was statistically significantly higher than when students were told that there were low consequences associated with the assessment.

Multi-Method Teacher Evaluation: Integrating Scores From Multiple Observational Tools
Ryan Kettler, Rutgers, The State University of New Jersey, Alexander Kurz, Arizona State University, and Linda Reddy, Rutgers, The State University of New Jersey

Multi-method educator evaluation has received national attention and shaped statewide policy for promotion and dismissal practices. Limited research exists on how evaluation tools offer shared and unique contributions to measurement of effectiveness. The study applies video coding and a multi-trait, multi-method matrix to examine the relations among concurrent observational systems.

Student Growth Percentiles Based on MIRT: Implications of Calibrated Projection
Scott Monroe, Li Cai, and Kilchan Choi, CRESST

Student Growth Percentiles (Betebenner, 2009) are used to locate a student's current score in a conditional distribution based on the student's past scores. This research presents an alternative method based on multidimensional IRT and calibrated projection linking that accounts for measurement error. Simulated and empirical examples are provided.

Behaviorally Anchored Rating Scales: An Application for Evaluating Teaching Practice
Michelle Martin-Raugh, Richard J. Tannenbaum, Cynthia M. Tocci, and Clyde Reese, Educational Testing Service, Princeton, NJ

We report the development of a Behaviorally Anchored Rating Scale (BARS) for measuring observed teaching practice. We describe the development and benefits of this measure, and compare its measurement properties to those of the Framework for Teaching (Danielson, 2011) by having trained raters evaluate video-taped teacher performances using both methods.

Sunday April 19, 2015
12:25 PM - 1:55 PM, Grand Ballroom, 7th Floor, Coordinated Session, L4

Using Principled Assessment Frameworks to Guide Test Development and Validation

Session Chair: Matthew Burke, National Commission on Certification of Physician Assistants
Session Discussant: Kirk Becker, Pearson VUE

To expand the work being done to further incorporate Principled Assessment Frameworks in operational practice, five organizations/groups will discuss the progress made in retooling traditional test development and validation techniques. In particular, the use of construct maps and task models to address practical problems in testing will be discussed.

The Development of Knowledge Requirement Scales in the Health Professions
Mark Raymond, National Board of Medical Examiners and Nance Cavallin, American Registry of Radiologic Technologists

Construct maps for 33 KSA domains in radiologic technology were developed using a methodology similar to that used for creating behaviorally-anchored rating scales. The 33 maps were formatted into a survey and field tested with educators and hiring managers who rated the level of knowledge expected of minimally-competent radiographers.

Developing a Skills Hierarchy Meaningful to Professionals and Test Professionals
Joshua Stopek, Henrietta Eve, and Matthew Schultz, American Institute of Certified Public Accountants

Implementing principled assessment frameworks (PAFs) requires changing the way that subject matter experts think about the nature of their construct. This paper focuses on the practical impact of PAFs on the development of a practice analysis survey, work done to validate its operational effectiveness, and improvements to ensure its success.

The Effects of Varying Grain Size on the Exchangeability of Item Isomorphs
Mary Ann Simpson, Audra Kosh, Lisa Bickel, Jeffrey Elmore, Eleanor Sanford-Moore, Heather Koons, and Marelle Enoch-Marx, MetaMetrics, Inc.

Grain size of cognitive models for automatically generated multi-digit addition and subtraction items was manipulated. 306 students responded to 24 generated items in fine, medium, or large grain size and item difficulties for the two forms within each grain size compared. Correlations were all above .80. Additional analyses reported.

Development and Empirical Validation of Assessment Engineering Task Model Grammars for Mathematics and General Science Assessments"
Richard M. Luecht, The University of North Carolina at Greensboro

This paper discusses task modeling research for the ASVAB Arithmetic Reasoning and General Science subtests. Expert-generated narratives and reverse-engineering techniques were employed to develop task model grammars (TMGs) for constructing effective cognitive task models. Empirical results demonstrate the recovery of item difficulties from TMG components. Practical implications are discussed.

Incorporating Construct Maps and Task Models in Standard Setting

Robert Furter, American Board of Pediatrics; Matthew J. Burke, National Commission on Certification of Physician Assistants; Deanna Morgan, and Pamela Kaliski, The College Board

This paper investigates the use of task models in the context of a modified Angoff standard setting procedure. Discussion will focus on the development and use of task models to aid panelists in the process of linking performance on items to Achievement Level Descriptors for an established score scale.

Sunday April 19, 2015
12:25 PM - 1:55 PM, King Arthur, 3rd Floor, Paper Session, L5

Innovations in Operational Testing

Session Chair: Han Yi Kim, Measured Progress
Session Discussant: Patrick Meyer, University of Virginia

Eliminating the Fourth Option in Multiple-Choice Items
Erkan Atalmis, Sutcu Imam University, Turkey, and Marianne Perie, Center for Educational Testing & Evaluation

Prior research has shown that three-option multiple choice items are just as reliable as four-option MCIs and offer many benefits. Different methods for eliminating one option from four option MCIs to construct three option MCIs could result in different item and test characteristics, as documented using a non-parametric IRT model.

Beyond Dichotomous: A Method for Collecting Better Test Data
Paul Curran, Kenyon College, Gambier, OH

Simple multiple choice test items are unable to collect more than correct/incorrect information without employing labor intensive coding of alternatives or complex psychometric methods that rely on large samples. This paper details a new take on old research allowing test-takers to provide forms of confidence in answers for partial credit.

Innovative Item Invariance Within the Context of Item Position Change
Ryan Glaze, Jenna Copella, and Leah Kaira, Pearson

Equating research has demonstrated that item parameter estimates for multiple-choice items vary as a function of item position. This study, using data from an end-of-course nursing exam, focuses on the extent to which the inclusion of innovative items may exacerbate the impact of item position changes and threaten parameter invariance.

Comparing Conversation-Based Scenarios to Traditional Assessment Methods
Haiying Li, University of Memphis, G. Tanner Jackson, and Diego Zapata-Rivera, Educational Testing Service

This paper examines outcomes from conversation-based tasks and compares them to traditional assessment methods. Three versions of the same assessment were developed to include multiple-choice only, multiple-choice with constructed response, and trialogue conversations (i.e., discussions between the test-taker and 2 virtual agents). Results will focus on students' learning and engagement.

Towards a Pro-Choice Strategy for Standardized Testing
Steven Culpepper and James Balamuta, University of Illinois at Urbana-Champaign, Champaign, IL

Prior research considered allowing examinees the flexibility to choose among exam items. New theoretical results demonstrate that choice increases item information if examinees have incentives to practice success arbitrage, which is defined as decisions that maximize expected utility. Experimental results support the psychometric value of choice when examinees have incentives.

Sunday April 19, 2015
12:25 PM - 1:55 PM, Renaissance Ballroom, 5th Floor, Paper Session, L6

DIF With Special Item Types

Session Chair: Ze Wang, University of Missouri
Session Discussant: Jessalyn Smith, CTB

Semiparametric Modeling and Testing of Differential Item Functioning (DIF)
Yang Liu, Brooke Magnus, and David Thissen, University of North Carolina, Chapel Hill, NC

Methods for modeling and testing DIF along a continuous covariate are often restrictive in parametric form. We develop a semiparametric approach that simultaneously estimates covariate-conditional item characteristic functions (ICFs) for non-anchor items, unconditional ICFs for anchor items, and conditional latent density distributions. A permutation DIF test is also proposed.

A Framework for Evaluating Score Comparability
Andrew Mroch, Dongmei Li, and Tony Thompson, ACT, Iowa City, IA

We propose a framework for evaluating comparability. We then illustrate the application of this framework to a mode comparability study. We recommend the framework for practitioners evaluating comparability and conclude that for the illustrative study data it was not reasonable to assume score comparability prior to applying statistical linking.

Using Multiple Linear Regression for DIF Assessment in Continuous Items
Hsiu-Yi Chao and Shu-Ying Chen, National Chung Cheng University, Chiayi County, Taiwan

This study proposed the multiple linear regression method (MLR) for DIF assessment in continuous items. Results of simulation study showed that MLR performed well in detection of uniform DIF. However, a further study to improve the performance of MLR in assessing nonuniform DIF is needed.

An Optimal Method for Detecting Item Drift
Jerome Clauser, American Board of Internal Medicine, Philadelphia, PA

Many methods exist for assessing the magnitude of item parameter drift. Unfortunately these methods do not provide a criterion for excluding items from the anchor. This paper uses error in examinee ability estimates as a criterion for evaluating anchors. A method for optimizing this criterion is presented.

Sunday April 19, 2015
12:25 PM - 1:55 PM, Seville Ballroom West, Lobby Level, Paper Session, L7

Comparing Equating Methods

Session Chair: Xia Mao, Pearson
Session Discussant: Lisa Keller, University of Massachusetts

The Impact of Different Anchor Test Structures in IRT Equating
Aolin Xie and Tammy Trierweiler, Prometric, Baltimore

Characteristics of anchor item sets were evaluated under the common item to an equated pool design. Anchor test length, the anchor test structure, and the underlying ability distribution of the equated new form were evaluated. Simulated data and data from a high stakes assessment were used to evaluate study conditions.

Using Robust Scale Transformation Methods for Multiple Outlying Common Items
Yong He, Zhongmin Cui, ACT, Inc. and Steven Osterlind, University of Missouri

Detection and elimination of outliers in common items, although improve equating accuracy, may cause inadequate content representation. Robust scale transformation methods have recently been proposed to solve this problem when only one outlier was present in a common item set. We applied the new approaches when multiple outliers are presented.

Evaluating Common Item Block Options When Faced With Practical Constraints
Amanda Wolkowitz and Susan Davis-Becker, Alpine Testing Solutions, Chesterfield, MO

This study evaluates the impact of common item characteristics on the outcome of equating in credentialing examinations when traditionally recommended representation is not possible. This research includes real data sets from multiple credentialing exams to test the impact of content representation, item statistics, and test length on equating results.

Evaluating Anchor-Item Designs for Concurrent Calibration With the GGUM
Seang-Hwane Joo, University of South Florida, Philseok Lee, University of South Florida and Jacob Seybert, Educational Testing Service

We developed MCMC procedures for concurrent calibration with the Generalized Graded Unfolding Model and conducted a Monte Carlo study to examine the efficacy of three anchor-item designs in vertical and horizontal linking scenarios. We will present these results and provide recommendations for pretest designs involving ideal point CAT applications.

Sunday April 19, 2015
12:25 PM - 1:55 PM, Toledo, 5th Floor, Paper Session, L8

Linking and Vertical Scaling

Session Chair: Francis Rick, University of Massachusetts Amherst
Session Discussant: Swaminathan Hariharan, University of Connecticut

Equating Task Complexity in Alternate Assessments
Fen Fan, University of Massachusetts Amherst and Louis Roussos, Measured Progress

A new alternate assessment design assesses students with items having different levels of complexity, which precludes standard statistical equating of the items. This study develops and demonstrates a new approach that enables the construction of a meaningful scale based on a combination of statistical and human judgment-based methods.

Scale Transformation in MIRT: A Parameter-Constraint Approach
Tsung-Han Ho, Jiyun Zu, and Lixiong Gu, Educational Testing Service, Princeton, NJ

There is a growing interest in MIRT application because modeling more than one dimension often improves the model fit. One limitation of applying MIRT in practice is the difficulty establishing equivalent scales of multiple traits. The performance of the parameter-constraint method is evaluated across various conditions.

Multidimensional IRT Scaling of AA-AAS Pilot Items Using Collateral Information
David R. King, Georgia Institute of Technology, Seung W. Choi, McGraw-Hill Education CTB, Anne Davidson, Smarter Balanced Assessment Consortium, and Sarah Hagge, Minnesota Department of Health

Alternate assessment items were scaled by fitting a confirmatory MIRT model to student and teacher responses. Teacher responses measured learner characteristics as common collateral information across forms and provided information for linking Math/ELA items and improving the measurement precision of Math/ELA ability scores.

Effect of Construct Shift on Vertical Scale Growth Estimates
Jane Rogers, Melissa Eastwood, and Hari Swaminathan, University of Connecticut

A simulation study was conducted to investigate the effect of construct shift on estimation of growth trajectories and to assess the feasibility of a bifactor model for measuring growth when the construct invariance assumption is violated. Growth trajectories were well-recovered with a bifactor model.

An Application of Multidimensional Vertical Scaling
Jonathan Weeks, Educational Testing Service, Princeton, NJ

This study evaluates design considerations in the development of a multidimensional vertical scale using empirical data from an assessment of reading component skills administered to examinees in grades 6 through 9. Three dimensional structures are considered: simple structure, bifactor, and unidimensional along with three linking methods.

Sunday April 19, 2015
12:25 PM - 1:55 PM, Valencia, Lobby Level, Paper Session, L9

Mixture IRT Models

Session Chair: Tim O'Neill, Pearson
Session Discussant: Pui-Wa Lei, Penn State University

The Confirmatory Mixture IRT Model for Inattentive Responses

Kuan-Yu Jin, The Hong Kong Institute of Education and Hui-Fang Chen, City University of Hong Kong

Inattentive responses, which can threaten measurement quality, are commonly observed in rating or Likert-type scale data. A mixture item response model was proposed to identify different kinds of inattentive responses. Simulation studies were conducted to evaluate the effectiveness of the new model.

A Random Item Mixture Nominal Response Model

Hye-Jeong Choi, Allan Cohen, University of Georgia and Brian Bottge, University of Kentucky

This study describes a random item mixture nominal response model. An important benefit of this model is that one can include covariates on persons and on individual response categories. A simulation study and a real data example will be provided to evaluate the usefulness of the model.

Guessing Detection Using Hybrid Mixture IRT Model With Response Times

Tongyun Li, Educational Testing Service and Hong Jiao, University of Maryland-College Park, College Park, MD

The present study investigates a hybrid mixture item response theory (IRT) model with response times (RTs) as covariates for the detection of guessing behavior. A simulation study is proposed to compare this approach with the mixture Rasch model (MRM)-RT approach (Meyer, 2008), which utilizes RTs in a log-normal function.

Sunday April 19, 2015
2:15 PM - 3:45 PM, Empire Ballroom, 7th Floor, Coordinated Session, M1

Psychometric Considerations for PARCC Assessments: Research From the Field Test

Session Chair: Enis Dogan, PARCC Inc.
Session Discussant: Michael Kolen, University of Iowa

PARCC is a state-led consortium working to more accurately measure student progress toward college and career readiness, developing a complex assessment system that will include an array of new item types. This session presents selected PARCC field test research studies to support the reliability and validity of test score interpretations.

Psychometric Research Studies and Data Collection Design
Bradley Moulder, Shameem Gaj, Kevin Meara, Jianbin Fu, Carolyn Wentzel, and Jing Miao, Educational Testing Service

The innovative design of the PARCC items and the structure of the PARCC assessment challenge traditional psychometric methods. This presentation provides an overview of the field test psychometric research, as pertains to scoring and scaling, the associated field test design consideration for collection of analysis data, and results to date.

Study of Device Comparability Within the PARCC Field Test
Leslie Keng, Laurie Laughlin Davis, Malena McBride, and Ryan Glaze, Pearson

PARCC's ultimate goal is to delivery its assessments using the widest variety of digital devices. A comparability study was conducted to compare student performance on computers and touch-screen tablets during the spring 2014 field-test administration. Implications of study results to psychometric work, item development and policies decisions will be discussed.

Mode Comparability
Terran Brown, Usama Ali, Jianshen Chen, Guangming Ling, Bradley Moulder, and Gautam Puhan, Educational Testing Service

PARCC computer-based (CBT) and paper-based (PBT) test forms are designed to be parallel, with the exception of technology-dependent items in the CBT forms. In the interest of fairness, this study examines the comparability of CBT and PBT item- and test-level scores, and the feasibility of scaling to support score comparisons.

Comparability of High School Mathematics End-of-Course Assessments
Kyunghee Suh, Lora Monfils, Igor Himelfarb, and Marna Golub-Smith, Educational Testing Service

PARCC offers mathematics end-of-course assessments for two pathways, specifically the Traditional and Integrated course sequences. This study examines the extent to which items for the two pathways can be placed on the same scale to support comparisons of student achievement in the domains assessed in the respective course sequences.

Sunday April 19, 2015
2:15 PM - 3:45 PM, Exchange, 11th Floor, Coordinated Session, M2

Theory-Based Item Generation for Mathematics Assessment and Instruction

Session Chair: Mary Ann Simpson, Lexile
Session Discussant: Linda Plattner, Illustrative Mathematics

Presenters: Mary Ann Simpson, Lisa Bickel, Jack Stenner, Jeff Elmore, Ellie Sanford-Moore, MetaMetrics, Inc.; William Fisher, University of California, Berkeley; Ruth Price, Lela Durakovic, Sandra Totten, Mark Kellogg, David Lines, and Donald Burdick, MetaMetrics, Inc.

A hybrid approach to automatic item generation combines the benefits of theory with the productivity offered by item templates. Our team will present papers on development of a strong theory of task difficulty in K-12 mathematics, identification of key task features, development of item generation software, and validation.

A Unified Theory of Task Difficulty In K-12 Mathematics
Jackson Stenner, Mary Ann Simpson, MetaMetrics, Inc.; William F. Fisher, University of California, Berkeley; and Donald Burdick, MetaMetrics, Inc.

Family Group Generation Theory - Large Scale Implementation
Lisa Bickel, Mary Ann Simpson, Ellie Sanford-Moore, Ruth Price, Lela Durakovic, and Sandra Totten, MetaMetrics, Inc.

Creating a Hybrid Math Item Generator
Mark Kellogg, Ryan Leathers, David Lines, and Lisa Bickel, MetaMetrics, Inc.

Initial Validation of Theory and Creation of Item Families
Mary Ann Simpson, Jeff Elmore, and Lisa Bickel, MetaMetrics, Inc.

Sunday April 19, 2015
2:15 PM - 3:45 PM, Grand Ballroom, 7th Floor, Invited Session, M3

Debate: Equal Interval Scales in Educational Testing: Attainable Goal or Myth?

Session Chair: Terry Ackerman, University of North Carolina Greensboro
Presenters: Wim van der Linden, CTB and Derek Briggs, University of Colorado

Ever since S.S. Stevens (1946) introduced his idea of different levels of measurement, the importance of these distinctions has been a matter of some controversy. In this debate, Derek Briggs will take the position that if measurement with clearly defined reference units is desirable of educational testing, then it is necessary for psychometricians to establish empirical criteria that must be met in support of such claims. In contrast, Wim van der Linden will argue that the ideal is a myth we should leave behind us and our current models already provide measurements that are extremely practical and lend themselves to fruitful theory building.

Sunday April 19, 2015
2:15 PM - 3:45 PM, King Arthur, 3rd Floor, Paper Session, M4

Subscore Recovery

Session Chair: Jonathan Beard, College Board
Session Discussant: Howard Everson, City University of New York

When Can We Improve Subscores by Making Them Shorter?
Richard Feinberg and Howard Wainer, National Board of Medical Examiners, Philadelphia

Subscores can be of diagnostic value for tests that cover multiple underlying traits. Some tests that contain items requiring ability spanning multiple traits may include such items on multiple subscores. In this study we show that the value of a subscore is always improved through the removal of such items.

Reporting Subscores in CAT Using a Bottom-Up Approach
Jing-Ru Xu, Michigan State University; Frank Rijmen, Seung W. Choi, and Sandip Sinharay, Pacific Metrics

This paper focuses on the implementation of the bottom-up approach. Two multidimensional models were analyzed with four dimensions. Three different MCAT approaches were compared using two ways of selecting subdomains under three different correlation designs. It shed lights on innovative methods of subscore reporting in CAT using different multidimensional models.

Subscores Aren't for Everyone: Alternative Strategies for Evaluating Subscore Utility
Mark Raymond and Richard Feinberg, National Board of Medical Examiners, Philadelphia, PA

Recent methods to substantiate subscore utility apply to all examinees in a sample and overlook individual differences. We propose an alternative method that determines whether subscores are useful for some examinees. The method, which draws on multivariate G-Theory, is sensitive to individual differences in profile variability and measurement error.

Examining Subscore Invariance Under Multidimensional Item Response Models
Jiyun Zu, Ho Tsung-Han, and Lixiong Gu, Educational Testing Service, Princeton

Multidimensional item response theory (MIRT) models have been shown to provide accurate and reliable subscores. A test fairness procedure of examining group invariance of MIRT-based subscores is described and demonstrated using data from two tests.

Sunday April 19, 2015
2:15 PM - 3:45 PM, Renaissance Ballroom, 5th Floor, Paper Session, M5

Reliability Related New Models

Session Chair: Brett Foley, Alpine Solutions
Session Discussant: William Skorupski, Kansas University

Decision Consistency and Accuracy for Multidimensional Models
Lee LaFond, Measured Progress and Won-Chan Lee, University of Iowa

This study developed a new procedure for estimating decision consistency and accuracy indices (DCA) using the bifactor and testlet response theory (TRT) models in tests employing testlets. The impact of placement of cut-scores and degree of multidimensionality on estimates of DCA indices between UIRT, TRT, and bifactor models was investigated.

A G-Theory Framework for Estimating Reliability Applied to Web-Based Measurement
Walter Vispoel and Murat Kilinc, University of Iowa, Iowa City, IA

A G-theory approach to estimating equivalence, stability, and equivalence and stability together is described and applied to 22 web-based measures of personality, self-concept, and social desirable responding. Results indicated that equivalence and stability indices analogous to Alpha and test-retest coefficients routinely overestimated reliability and sometimes markedly so.

Estimating Reliability by Bayesian Structural Equation Modeling
Wei Tang, Qi Guo, and Ying Cui, University of Alberta

Structural Equation Modeling Estimate of Reliability (SEMR) is one of the currently recommended methods to estimate reliability. The goal of the study is to compare two of SEMR's estimators: Bayesian estimator and maximum likelihood, under different simulation conditions. We are interested in their performance in small sample and non-normal conditions.

The Robustness of Structural Equation Modeling Estimates of Reliability
Qi Guo, Wei Tang, and Ying Cui, University of Alberta, Edmonton, Canada

The estimate of reliability using structural equation modeling tends to be biased if the analysis model is incorrectly specified. However, to correctly specify the analysis model is difficult in practice. Thus, this study is designed to explore the robustness of SEM estimates of reliability under misspecified models.

Sunday April 19, 2015
2:15 PM - 3:45 PM, Seville Ballroom East, Lobby Level,
Coordinated Session, M6

Psychometric Considerations in Linking to Survey Assessments

Session Chair: Rochelle Michel, Educational Testing Service
Session Discussant: Kadriye Ercikan, The University of British Columbia

Presenters: Meng Wu, Xueli Xu, Rochelle Michel, and Yue Jia, Educational Testing Service

This coordinated session consists of four papers that will highlight best practices in linking to and between survey assessments. The papers will include a general description of psychometric considerations in linking studies, as well as expositions of the various challenges and opportunities in linking to and between survey assessments.

Sunday April 19, 2015
2:15 PM - 3:45 PM, Seville Ballroom West, Lobby Level, Paper Session, M7

Score Reporting
Session Chair: Joshua Goodman, Pacific Metrics
Session Discussant: April Zenisky, University of Massachusetts Amherst

Exploring the Effectiveness of a Measurement Error Tutorial for Teachers
Diego Zapata-Rivera, Rebecca Zwick, and Margaret Vezzu, Educational Testing Service, Princeton, NJ

The goal of this study was to explore the effectiveness of a focused web-based tutorial in helping teachers to better understand confidence bands in test score reports. Results showed a significant difference in comprehension scores between groups that received a tutorial and the control group (no tutorial).

Investigating Test User Interpretations and Actions Based on Score Reports
Timothy O'Leary, John Hattie, and Patrick Griffin, Melbourne University, Melbourne, Australia

Score reports and the appropriateness of the interpretations that users make play an integral part in test validity. This paper provides a rationale for research focused on validating interpretations and actions and presents findings from the first of a series of studies relating to user interpretations and actions.

Designing Alternate Assessment Score Reports That Maximize Instructional Impact
Sheila Wells-Moreaux, Amy K. Clark, Gretchen Anderson, Meagan Karvonen, and Neal Kingston, University of Kansas

This paper describes the process for designing alternate assessment score reports to be actionable for teachers and parents. We used focus groups methods to refine prototypes, examine teachers' use of scores to plan instruction for students with significant cognitive disabilities, and evaluate the utility of reports to guide instructional decision-making.

The Relative Performance Index: Defusing Simpson's Paradox
Kyle Nickodem, Ernest Davenport, Jr., Gareth Phillips, University of Minnesota; Edmund Graham, University of Illinois at Urbana-Champaign; and Deanna Garcia, Press Ganey Associates

Instead of relying on a potentially misleading aggregate mean score, the Relative Performance Index provides policy makers and the public with a single summary value for assessment data that adjusts for Simpson's Paradox and gives voice to disaggregated results.

Sunday April 19, 2015
2:15 PM - 3:45 PM, Toledo, 5th Floor, Coordinated Session, M8

Thinking About Validity in Measuring Teacher and School Effectiveness

Session Chair: Burcu Kaniskan, NCBE
Session Discussant: Michael Kane, Educational Testing Service

Numerous studies examine measurement of teacher/school effectiveness. Yet, these studies include several methodological and practical issues which constitute threats to validity (Braun, Chudowsky, & Koenig, 2010). This session addresses these limitations and provides innovative solutions in measuring teacher and school effectiveness.

Impact of Ignoring in Cross-Classified Multiple Membership Data Structures
Dan Murphy, Burcu Kaniskan and Ahmet Turhan, Pearson and NCBE

This study compared the use of a three-level growth-curve model with that of a cross-classified growth curve model and a cross-classified multiple membership growth-curve model for handling cross-classified multiple membership data structures. Results indicate different models would lead to different conclusions about the nature of teacher effects on student growth.

Assessment of Growth: A Comparison of Models for Projecting Growth
Burcu Kaniskan, Swaminathan Hariharan, and Jane Rogers, NCBE and University of Connecticut

The purpose of this study is to compare student growth models with respect to their ability to: (a) accurately project students math scores at a future point; and (b) correctly classify students' proficiency levels at future point in time. Nine growth models derived from four different methods are compared.

Using VAM to Triage Teacher Evaluation
Kathy McKnight, Dan Murphy, and Doug Harris, Pearson and Tulane University

The valid use of VAM as a measure of teacher effectiveness has been in question for some time (McCaffrey, et al., 2003). We describe the design of a triage-like evaluation process, incorporating VAM instead as a screener, to flag extreme (high and low) performance and to triangulate evaluation data.

Combining Measures of Teaching Effectiveness With the Hierarchical Rater Model
Jodi Casabianca, University of Texas at Austin

This research evaluates a variation of the hierarchical rater model for estimating teaching effectiveness based on multiple measures. The augmented parameterization may incorporate measures like classroom observation ratings and student outcomes. The models are fitted to Measures of Effective Teaching data to compare estimated traits and estimates from standard approaches.

Sunday April 19, 2015
2:15 PM - 3:45 PM, Valencia, Lobby Level, Coordinated Session, M9

Applications and Advances of Multidimensional IRT With Stochastic Approximation Methods

Session Chair: Lauren Harrell, University of California Los Angeles
Session Discussant: Sandip Sinharay, Pacific Metrics

Stochastic approximation methods, such as the Metropolis-Hastings Robins-Monro algorithm, have increased the flexibility of item response modeling and have broadened applications of item factor models. Advances in multidimensional and hierarchical IRT estimation through stochastic approximation are applied to large-scale assessment methodology, teacher evaluation, and rater-effects estimation in task-based performance assessments.

A Crossed Random Effect MIRT Model for Multiple Rater Data
Larry Thomas and Li Cai, University of California, Los Angeles

This research investigates a crossed random effect multidimensional IRT model to consistently and accurately score examinees on multiple-rated performance tasks and evaluate raters' use of scoring rubrics. Simulation results show that item parameters, examinee and rater estimates, and standard errors are accurately estimated via a stochastic approximation algorithm.

Examining NAEP Fourth Grade Mathematics Using a Multilevel Item Factor Analysis Model
Nathan Dadey, University of Colorado-Boulder

Empirical dimensionality, as defined through exploratory methods, often fails to align to test blueprints or other substantive theory. Relatedly, empirical dimensionality can vary by level of aggregation. This work examines these issues in the context of NAEP fourth-grade mathematics assessment through the application of a multilevel item factor analysis model.

Plausible Value Imputation From Multidimensional IRT Models Using Stochastic Approximation Methods
Lauren Harrell and Li Cai, University of California, Los Angeles

The Metropolis-Hastings Robbins-Monro algorithm is adapted to perform multidimensional IRT calibration simultaneously with latent regression. The properties of imputation of plausible values for large-scale assessments is examined using both simulation studies of complex MIRT and analysis of NAEP Science data under the framework-specified data generating model.

Multilevel Item Factor Analysis With Covariates for Teacher Evaluation Surveys
Megan Kuhfeld and Li Cai, University of California, Los Angeles

We compare latent score estimates from a multilevel multidimensional item factor model to traditional scoring methods using simulations and data from a teacher evaluation survey to examine score bias in teacher practice and class environment constructs when important student background covariates are included or excluded from the calibration model.

Sunday, April 19, 2015
4:00 PM–7:00 PM, Burnham Room, 8th Floor

NCME Board of Directors Meeting

Members of NCME are invited to attend as observers.

Sunday April 19, 2015
4:05 PM - 5:05 PM, Camelot, 3rd Floor, Electronic Board Session, Paper Session, N1

Electronic Board #1
Do the Step Parameters Matters When Doing DIF Detection on Dichotomized Responses?
Shuwen Tang, Wen Zeng, and Cindy Walker, University of Wisconsin-Milwaukee, Milwaukee, WI

This study focuses on the effect of having different step parameters in polytomous items and dichotomizing them, prior to testing for differential item functioning (DIF). A simulation study was conducted to compare the performance of SIBTEST and Poly-SIBTEST, in terms of the Type I error and power, under these conditions.

Electronic Board #2
Factors Influencing the LR for Detecting DIF in Within-Multidimensional Data
Hui-Fang Chen, City University of Hong Kong; Kuan-Yu Jin, and Wen-Chung Wang, Hong Kong Institute of Education

Simulations were conducted to investigate factors that influence the performance of the logistic regression method in assessing differential item functioning (DIF) for within-multidimensional data. Several factors were manipulated when non-uniform DIF exists. It was found that the matching variable significantly influenced the accurately detection rates under the bidirectional DIF pattern.

Electronic Board #4
Applying Wald Test to Detect Multi-group DIF in CDM
Likun Hou, Educational Testing Service and Jimmy de la Torre, Rutgers University

Previous studies showed Wald test is effective to detect DIF between two groups in cognitive diagnosis models. This simulation study extends the use of Wald test to detect DIF for more than two groups in the context of the DINA model. Results show Wald test is promising for this purpose.

Electronic Board #5
Differential Item Functioning for Testlet-Based Tests in Multilevel Data
Jin Zhang and Changhui Zhang, ACT, Iowa City

This study uses a logistic mixed model to detect DIF for testlet-based tests with multilevel student data. A simulation study is conducted to compare the performance of the logistic mixed model and Mantel-Haenszel test under various degrees of DIF, local item dependence, and local person dependence.

Electronic Board #6
DIF Detection and Prediction With Different Focal Group Designations
Jessica Loughran and Neal Kingston, University of Kansas, Lawrence

This study examined the effect of different focal group designations on: 1) the detection of DIF, and 2) the prediction of DIF from individual item features. Results indicated that different focal group designations influence both DIF detection and prediction. This research has implications for test development and validity studies.

Electronic Board #7

Application of Differential Person Functioning for the Scale Comparability

Dong Gi Seo, National Registry of Emergency Medical Technicians, Columbus, OH

Differential person functioning (DPF) can be applied to investigate the comparability of examinees within different subgroup with small sample sizes, and few or even no common items. This study shows that DPF is appropriate for the purpose of scale comparability regardless of the size of the subgroups and common items.

Electronic Board #8

Detecting DIF in Multidimensional Data Using MIMIC and IRT-LR-DIF

Okan Bulut, University of Alberta; Soo Lee, and Youngsuk Suh, Rutgers, The State University of New Jersey

This study introduces the multidimensional forms of the MIMIC-interaction model and the IRT-LR-DIF test for detecting DIF in multidimensional data. The mechanisms of these two DIF approaches are presented and the performances of the two methods are compared via a Monte Carlo simulation study under various simulation conditions.

Electronic Board #9

Using DIF to Investigate Reading Proficiency's Impact on Science Achievement

Pierre Brochu, University of Toronto/OISE, Whitby, Canada

This study investigates the impact of reading proficiency on science achievement using DIF in a multidimensional model. Canadian Grade 8 students responded to a national assessment in either English or French and results in science are compared across language groups controlling for reading achievement.

Electronic Board #10

Assessing the Latent Dimensional Structure of a Cross-Cultural Data

Minhee Seo, KICE, Yun Seok Choi, Korea National University of Transportation, and Kyong Hee Chon, Kangnam University

We examined the latent dimensional structure of a cross-cultural data both in math cognitive and math non-cognitive areas. The results indicated that the dimensional structure and item characteristics of math achievement test should not be applied in the same way to those of non-cognitive scale without consideration of cultural differences.

Electronic Board #11

Explanatory Models for Understanding Differential Item Functioning

Jennifer Brussow, William Skorupski, and Jessica Loughran, University of Kansas, Lawrence, KS

This study evaluates the use of specific item-level features as explanatory variables for understanding DIF. Two methods of DIF identification/explanation are considered: 1) two-stage DIF + regression, and 2) a simultaneous, hierarchical approach. Data are simulated across three grade levels to determine comparability of results across non-linked assessments.

Electronic Board #12
A Comparative Study of DIF Procedures for Computerized Adaptive Tests
Joseph Rios, Educational Testing Service, Princeton, NJ

The objective of this study is to introduce a new CAT-DIF procedure referred to as the fixed-effects likelihood ratio test (FE-LRT) and to compare its performance with the logistic regression, CATSIB, and CATSIB-E methods. Overall, the FE-LRT demonstrated superior type I error, power, and correct DIF type identification rates.

Electronic Board #13
The Influence of DIF/DBF on Ability Estimation
Kevin Cappaert, University of Wisconsin - Milwaukee, Milwaukee, WI

The factors test length, bundle size, uniform DIF/DBF, non-uniform DIF/DBF, and reference/ focal group balance were investigated to determine their influence on ability estimation. Results indicate uniform DIF/DBF has an influence on ability estimation bias while non-uniform DIF/DBF was found to influence RMSE and the standard error of ability estimates.

Electronic Board #14
Measurement Invariance Assessment Under Cognitive Diagnostic Models
Xiaomin Li and Wen-Chung Wang, The Hong Kong Institute of Education, Hong Kong

In CDMs, all existing methods for DIF detection treated the latent binary attributes as qualitatively identical even when items exhibit DIF, which contradicts the common knowledge of DIF and thus misleading. In this study, we acknowledge such fallacy and propose a new method for DIF assessment in CDMs.

Electronic Board #15
Hierarchical Linear Modeling to Examine Comparability of Testing Modes
Heather Rickels, Wei Cheng Liu, Stephen Dunbar and Catherine Welch, Iowa Testing Programs, University of Iowa, Coralville, IA

This study examined the use of hierarchical linear modeling (HLM) when investigating comparability of computer-based testing (CBT) and paper-and-pencil testing (PPT). Specifically, the HLM approach was used to examine potential clustering effects and identify outlying schools. Failing to examine clustering effects may lead to erroneous conclusions of significance or non-significance.

Electronic Board #16
DIF Analysis for Short Tests With Multilevel Data
Ying Jin and Christian Flack, Middle Tennessee State University

Four DIF methods were compared to investigate their abilities to account for low reliability and multilevel data structure when mean ability difference was present. The results showed that DIF methods with latent covariates outperformed DIF methods with observed covariates under conditions simulated in the current study.

Electronic Board #17

The Effects of Discrimination Parameters on DIF Detection Methods

Nathan D. Minchen and Youngsuk Suh, Rutgers, The State University of New Jersey, New Brunswick, NJ

This study examines the effect of the discrimination parameters of anchor items on the performance of several popular Differential Item Functioning (DIF) detection methods, and further explores the performance of the Combined Decision Rule DIF test (Penfield, 2003).

Electronic Board #18

Assessment of DIF In Testlet-Based Items Using 2pl-Multilevel Measurement Model

Wei Xu and David Miller, University of Florida

In this study, DIF analysis among testlet-based items is conducted by the 2PL-multilevel measurement model (extended from the model developed by Beretvas and Walker (2012)). The proposed study will further our understanding about DIF detection and inform practitioners with regards to the choice of testlet-based items.

Sunday April 19, 2015
4:05 PM - 6:05 PM, Empire Ballroom, 7th Floor, Paper Session, N2

Equating: Small Samples and Testlets

Session Chair: Thomas Proctor, College Board
Session Discussant: Michael Jodoin, NBME

Bifactor MIRT Observed-Score Equating for Testlet-Based Tests
Juan Chen, National Conference of Bar Examiners and Wei Tao, ACT, Inc.

Three MIRT observed-score equating methods for testlet-based tests are compared using the bifactor model. Method I applies the unidimensional approximation of composite parameters; Method II requires a parameter reduction process through integration; Method III uses the full-information MIRT procedure. The equipercentile method is used as the baseline for comparison.

Minimum Sample Size Requirements for Equating of Mixed-Format Tests
Ja Young Kim, Ja Young Kim, Wei Tao, and YoungWoo Cho, ACT, Inc.

Very few studies have investigated the sample size requirements that lead to acceptable equating results with mixed-format tests. This study will estimate minimum sample sizes considering test length, scoring weights on CR and correlation between multiple choice (MC) and constructed response (CR) sections in mixed-format tests using equipercentile equating method.

The SiGNET Model for Small Sample Equating: A Practical Application
Xuan (Adele) Tan and Gautam Puhan, Educational Testing Service, Princeton, NJ

This study will evaluate an improved data collection design, the Single Group Nearly Equivalent Test (SiGNET) design, in dealing with small sample equatings using data from tests assembled with this design. Different sample sizes and equating methods will be examined. Preliminary results showed promise over traditional common item equating design.

Investigating Three IRT Perspectives on Equating Testlet-Based Tests
Mengyao Zhang, Hyung Jin Kim, Kyung Yong Kim, Won-Chan Lee, Euijin Lim, Shichao Wang, and Robert Brennan, The University of Iowa, Iowa City

The purpose of this study is to empirically investigate the performance of three different IRT perspectives on equating testlet-based tests. Both IRT true score and observed score equating are performed. Varying degrees of local item dependence and other test and sample characteristics are considered.

Sunday April 19, 2015
4:05 PM - 6:05 PM, Grand Ballroom, 7th Floor, Paper Session, N3

CAT: Test Generation

Session Chair: Anna Topczewski, Pearson
Session Discussant: John Willse, University of North Carolina Greensboro

Evaluating Statistical Targets for Assembling Parallel Mixed-Format Test Forms
Dries Debeer, University of Leuven, Usama Ali, and Peter van Rijn, Educational Testing Service

Mixed-format test are widely used in practice because of their increased validity. Because in many large-scale assessments parallel forms are needed, the theoretical framework to generate the test forms is critical. This project investigates the performance of different statistical targets in the linear assembly of mixed-format tests.

A Methodology for Multilingual Automatic Item Generation
Keith Boughton, McGraw-Hill Education CTB; Mark Gierl, Changhua Rich, University of Alberta; and Lorena Houston, McGraw-Hill Education CTB

AIG is a method in which various procedures found in educational fields are used to create large numbers of items. Tests are also administered in different languages, which require testing programs to accommodate multilingual testing. This research will employ a new method for multilingual automatic item generation.

A Novel Approach to Quantify Semantics of Automatically Generated Items
Syed Muhammad Fahad Latifi, Mark Gierl, Ren Wang, and Andong Wang, University of Alberta, Edmonton, Canada

We present a novel unsupervised approach that extends the Compositional Distributional Semantic Model (CDSM) to measure the semantic relatedness among the pool of automatically generated items. Generated items from the medical science domain were used. We found our measure sensitive in indexing the semantic-heterogeneity of item pools.

Consideration of Item Position Effects in Computerized Adaptive Testing
Raphael Bernhardt, Andreas Frey, and Sebastian Born, Jena University, Jena, Germany

A procedure to account for item position effects in computerized adaptive testing (CAT) is proposed. The application of the procedure to empirical data (N=1,632) revealed general item position effects but no item-specific effects. The model estimates can be used in CAT to avoid suboptimal item selection and biased ability estimates.

Optimal Reassembly of Shadow Tests in CAT
Seung W. Choi, Jie Li, Karin Moellering, and Wim van der Linden, McGraw-Hill Education CTB, Monterey

Even in the age of abundant and fast computing resources, security and concurrency requirements still put an uninterrupted delivery of computer-adaptive tests at risk. For the particularly compute-intensive shadow test approach to CAT we therefore examine strategies reducing the number of (concurrently) reassembled shadow tests without compromising its measurement quality.

Sunday April 19, 2015
4:05 PM - 6:05 PM, Renaissance Ballroom, 5th Floor, Paper Session, N4

DCM & Diagnostic Models

Session Chair: Liru Zhang, Delaware Department of Education
Session Discussant: Jonathan Templin, University of Kansas

A Generalized Approach to Defining Item Discrimination for DCMs
Robert Henson, University of North Carolina at Greensboro, Lou DiBello, University of Illinois at Chicago, and Bill Stout, University of Illinois at Champaign-Urbana,

One advantage of a DCM approach is that characteristics of the items can be used to refine an assessment. This paper discusses a unified approach for identifying good items that generalizes to multinomial models (i.e., several responses/item modeled, not just right/wrong) for diagnostic classification models.

Fitting Diagnostic Classification Models to Distractor-Driven Tests for Validation Purposes
Benjamin R. Shear, Stanford University; Louis Roussos, Measured Progress; Louis DiBello, and William Stout, Learning Sciences Research Institute, University of Illinois at Chicago, Robert Henson, University of North Carolina at Greensboro

This paper fits a new generalized diagnostic classification model for option-based scoring to a distractor-driven test of student misconceptions in middle school geometry. Q matrix specification, model fit statistics and examinee classifications provide validity evidence by testing the hypothesized cognitive structure of responses and informing scoring procedures.

The Effects of Mixture-induced Local Dependence on Diagnostic Classification
Thomas McCoy and John Willse, University of North Carolina Greensboro, Greensboro, NC

Diagnostic classification models assume local independence (LI) for performing classification into skill mastery profiles. However, the impact of violating LI has not been well studied. A simulation study on classification rates is presented after introducing systematic within-class variation from mixtures.

A Strategy-Based Diagnostic Classification Model
Ning Yan, Independent Consultant; Yuehmei Chien, and Chingwei Shin, Pearson

We present a new diagnostic classification model designed to support the modeling of multiple response strategies per test item, where each strategy utilizes a different subset of the latent attributes under assessment. The model has an intuitive formulation like DINA, but is closely related to more general frameworks like LCDM.

Item Fit Evaluation in Cognitive Diagnosis Modeling
Miguel Sorrel, Francisco Abad, Julio Olea, Universidad Autónoma de Madrid; and Juan Barrada, Universidad de Zaragoza, Madrid, Spain

In the field of cognitive diagnosis modeling, there has been scarce research related to the item fit evaluation. Based on a simulation study, this study investigate the performance of various item fit statistics (Wald statistic, RMSEA, and S-X2) and provide information about usefulness of these indexes on different scenarios.

Longitudinal Cognitive Diagnosis Model Application of Latent Transition Analysis
Yasemin Kaya and Walter Leite, University of Florida, Gainesville, FL

The purpose of our study was to develop a longitudinal model for cognitive diagnosis models, which will be applied to repeated measurements in order to monitor attribute stability of the individuals, and to account for respondent dependence. We developed a model based on latent transition analysis model in cognitive diagnosis.

Sunday April 19, 2015
4:05 PM - 6:05 PM, Seville Ballroom East, Lobby Level, Paper Session, N5

Applied International Assessment

Session Chair: James Ingrisone, Pearson
Session Discussant: Hua-Hua Chang, University of Illinois at Urbana-Champaign

Comparing the MSLQ Structure Models Between Korea and the U.S.
Jiyoung Yoon, Yoonsun Lee, Mi-jin Kwon, Misun Kim, and Hye-young Kang, Seoul Women's University

This study examined measurement invariance of the Motivated Strategies for Learning Questionnaire (MSLQ) between Korea and the U.S. The results showed the same structure between the two countries. Additionally, Multiple Indicators and Multiple Causes (MIMIC) model investigated the effect ethnical identification and gender on the structure of a two-factor model.

Affective Characteristics Predicting 15-Year-Old Students' Mathematical Literacy Skills in Turkey
Ergül Demir, Ankara University Educational Sciences Faculty, Ankara, Turkey

The aim of this study was to examine the affective characteristics of the 15-year-old students in Turkey, significantly predicting their mathematical literacy skills and competencies, according to the PISA 2012 results. According to modeling studies ongoing, it can be possible to determine a significant secondary-level structural equation modeling.

Socioeconomic Justice in Countries and Its Relation to Education
Mustafa Yilmaz, University of Kansas, Lawrence, KS

PISA is one of the leading large-scale assessments in comparative education. For the first time in 2012, PISA surveyed students for the private tutoring. By using their data this study examines what PISA supplies us about socioeconomic justice in countries and how does it relate to private supplementary educational services.

Understanding and Operationalizing Cultural Sensitivity in Assessment Practices
Edynn Sato, Pearson, San Francisco, CA

This presentation discusses the construct of cultural sensitivity. Research explicating its key elements will be presented, and an approach to operationalizing cultural sensitivity, so that we can ensure our assessment practices address the challenges of testing across different cultural groups, as well as yield valid, meaningful outcomes will be proposed.

An Error Analysis Examining International Assessments and Resulting Country Equivalence
John Poggio and Susan Gillmor, University of Kansas, Lawrence, KS

International assessments explore country standing, and the magnitude of differentiation among countries. This investigation studies if examinees make comparable errors regardless of score attained. Analyses show that errors within a country are similar regardless of scores attained, but error analyses reveal country differences suggesting that country instruction is not equivalent.

Sunday April 19, 2015
4:05 PM - 6:05 PM, Seville Ballroom West, Lobby Level,
Coordinated Session, N7

Exploiting Technology in the Service of Assessment for Learning

Session Chair: Cindy Ziker, SRI International
Session Discussant: Jim Minstrell, CADRE

Presenters: Caroline Wylie, Educational Testing Service, Cindy Ziker, SRI International, Margaret Heritage, WestEd and Kurt VanLehn, Arizona State University

This symposium describes the components of a formative assessment system for technology-enhanced classrooms. In order to fully leverage its power, one must integrate technology into a curriculum, and educators and students must know how to wisely exploit that technology for learning (Bennett & Gitomer, 2008; Heritage, 2007; OECD, 2005).

This session will present four topics relevant to exploiting technology in the service of formative assessment for learning:

Supporting Formative Assessment With Technology
Margaret Heritage, WestEd

Technology tools have not featured very prominently in classroom formative assessment practice. This paper will discuss some of the limitations of that technology tools have suffered from in the past with respect to formative assessment, and will describe the affordances of a number of current tools for effective formative assessment.

Validating and Using Learning Progressions to Support Mathematics Formative Assessment
Caroline Wylie, Malcolm Bauer and Meirav Arieli-Attali, Educational Testing Service

Learning progressions characterize conceptual thinking and understanding at increasing levels of sophistication. We use three LPs (Equality and Variable; Proportional Reasoning; Functions and Linear Functions) in a study focused on LPs to guide development of formative assessment tools and to support teachers' interpretation of student results against the progressions.

Dashboards in FACT!
Kurt VanLehn, Arizona State University

The Mathematics Assessment Project (http://map.mathshell.org/) has developed over 70 classroom Formative Assessment Lessons (FALs) that address the mathematical practices of the Common Core State Standards for Mathematics. The activities and software are being developed in a series of design experiments and usability studies. This talk describes the initial studies' findings.

New Protocols New Contexts: Observing Formative Practice in Technology-enhanced Classrooms
Cindy Ziker, Geneva Haertel, Harold Javitz and Terry Vendlinski, SRI International

Protocols must be sensitive to the technology genre employed (e.g., games, simulations, interactive computer tasks) and the type of data available. This presentation describes how and why existing observation protocols were leveraged to develop and test an observation protocol that describes teachers' formative practices when using a game-based curriculum.

Sunday April 19, 2015
4:05 PM - 6:05 PM, Toledo, 5th Floor, Coordinated Session, N8

Automated Scoring of Nontraditional Forms of Assessment

Session Chair: Frank E. Williams, Educational Testing Service
Session Discussant: Claudia Leacock, McGraw-Hill Education

Automated scoring of nontraditional measures present many challenges for quality measurement. Research is presented for tasks that assess the adequacy of financial statements with multiple solutions, assess clinical reasoning skills using medical notes, evaluate structured letters of recommendation for scholastic admissions, and score course papers for grading and feedback purposes.

Facilitating Automated Scoring for Simulated Balance Sheet Items
J. Stopek, AICPA

The more flexibility provided candidates, the more complicated it is to score their responses. In this paper, we describe and evaluate a rule-based scoring methodology that handles alternatively correct presentations in a fully automated way for spreadsheet-based items.

Automated Scoring of Patient Notes in a Medical Licensure Examination
S.G. Baldwin, P. Harik, B.E. Clauser, M. Winward, and P. Baldwin, National Board of Medical Examiners

Automated scoring was applied to the new version of the Patient Note (PN) component of the USMLE® Step 2® examination. The present study evaluates and compares the accuracy and efficiency of key features, extracted by an n-gram-based NLP algorithm, in predicting expert ratings of patient notes.

Automated Scoring of Text in Structured Letters of Recommendation
F.J. Breyer, F.E. Williams, M. Heilman, J. Blackmore, D. Klieger, and Laura Ridolfi-McCulla, Educational Testing Service

An automated scoring engine is applied to structured recommendation letters that consist of text provided by faculty for student applicants to higher education programs. The engine extracts features reflective of stereotypic or student-specific commentary, and positive versus non-positive sentiment. A scoring model is built and evaluated to predict admission status.

Automated Scoring of Writing Samples From Course Work
C. Ramineni, Educational Testing Service

Automated scoring is typically common for impromptu timed essay tasks. This paper describes the procedures, challenges, and results of attempting automated evaluation and scoring of longer untimed writing samples produced as part of course work compared to the automated evaluation of essay type writing samples written under timed conditions for a high-stakes test.

Sunday April 19, 2015
4:05 PM - 6:05 PM, Valencia, Lobby Level, Paper Session, N9

Comparing Standard Setting Methods

Session Chair: Jonathan Weeks, Educational Testing Service
Session Discussant: Steve Ferrara, Pearson

The Bookmark Method of Standard Setting: Issues and Research
Amin Saiar, PSI Services LLC, Alan Socha, American Registry for Diagnostic Medical Sonography, and John Weiner, PSI Services LLC

Establishing standards for examinee performance is a longstanding issue in testing for which there are numerous considerations and challenges. This presentation will explore key issues, empirical research findings, and considerations for the Bookmark method of standard setting in the context of criterion-referenced measurement (e.g., credentialing, licensure).

Effect of Content Knowledge on the Precision of Angoff Judgments
Melissa Margolis, Brian Clauser, Janet Mee, National Board of Medical Examiners; and Jerome Clauser, American Board of Internal Medicine

This study investigated whether correctly answering an item is a significant factor in predicting judge behavior. Judges first answered and then provided Angoff judgments for a set of 45 items. Judgments were significantly lower for items that judges answered incorrectly, suggesting that content knowledge has important implications for standard-setting outcomes.

Setting Standards in Adaptive Testing: An Adaptive Bookmark Method
Xin Luo, Michigan State University; Priya Kannan, and Richard Tannenbaum, Educational Testing Service

Traditional panel-based standard-setting procedures are not directly applicable to adaptive tests. We propose an adaptive Bookmark method (a-BM) as one solution. The a-BM reduces the number of items reviewed by panelists, while includes common items to facilitate discussions. Results of a simulation support the measurement quality of the a-BM approach.

Comparing Standard Setting Methods for a Likert Scale Test
Yoonsun Lee, Seungho Park, and Hyeyoung Kang, Seoul Women's University, Seoul, Republic of Korea

The purpose of this study is to compare standard setting method (Angoff, body of work, and empirical method) indifferent test response format (7-point, 6-point and 5-point scale) of a Likert scale test. The result revealed that the cut scores from nine combinations (three method X three formats) were differ

Participant Index

Participant Index

A

B

Participant Index

C

Participant Index

D

E

F

Participant Index

G

H

Participant Index

I

J

K

Participant Index

L

Participant Index

M

Participant Index

N

O

P

Q

Participant Index

R

S

Participant Index

T

V

Participant Index

W

X

Participant Index

Y

Z

Contact Information for Individual and Coordinated Sessions First Authors

Akbay, Lokman
Rutgers University
lokmanakbay@gmail.com

Alagoz, Cigdem
University of Georgia
cigdem@uga.edu

Albano, Anthony D.
University of Nebraska
albano@unl.edu

Ali, Usama
Educational Testing Service
uali@ets.org

Allalouf, Avi
National Institute for Testing and Evaluation
avi@nite.org.il

Ames, Allison J.
University of North Carolina Greensboro
ajames@uncg.edu

Anderson, Daniel
University of Oregon
daniela@uoregon.edu

Applegate, Greg
Pearson
gregory.m.applegate@gmail.com

Arce-Ferrer, Alvaro J.
Pearson
alvaro.arce-ferrer@pearson.com

Atalmis, Erkan
Sutcu Imam University, Turkey

Attali, Yigal
ETS
yattali@ets.org

Austin, Bruce W.
Washington State University
bruce.austin@email.wsu.edu

Ayers-Wright, Elizabeth
American Institutes for Research
eayers@air.org

Baldwin, Peter
National Board of Medical Examiners
pbaldwin@nbme.org

Banks, Kathleen
Middle Tennessee State University
kathleen.banks@mtsu.edu

Barrett, Matthew E.
The Georgia Institute of Technology
matthew.barrett@gatech.edu

Basaraba, Deni
Southern Methodist University
dbasaraba@smu.edu

Beaver, Jessica L.
Washington State University
jessica.l.beaver@email.wsu.edu

Béguin, Anton
Cito
anton.beguin@cito.nl

Bejar, Isaac
ETS
ibejar@ets.org

Bernhardt, Raphael
Jena University
raphael.bernhardt@uni-jena.de

Bertling, Jonas P.
ETS
jbertling@ets.org

Betebenner, Damian
Center for Assessment
dbetebenner@nciea.org

Betts, Joseph
Pearson Vue
jbetts5118@aol.com

Bezruczko, Nikolaus
Indiana University Health
nbezruczko@msn.com

Bickel, Lisa
MetaMetrics, Inc.
lbickel@lexile.com

Contact Information for Individual and Coordinated Sessions First Authors

Boughton, Keith
CTB/McGraw-Hill
keith_boughton@ctb.com

Bradshaw, Laine
The University of Georgia
laineb@uga.edu

Brennan, Robert
University of Iowa
robert-brennan@uiowa.edu

Broatch, Jennifer
Arizona State University
jennifer.broatch@asu.edu

Brochu, Pierre
Council of Ministers of Education (Canada)
p.brochu@cmec.ca

Brussow, Jennifer
University of Kansas
jbrussow@ku.edu

Bryant, William
ACT
bill.bryant@act.org

Buchholz, Janine
DIPF (German Institute for International Educational Research)
buchholz@dipf.de

Bukhari, Nurliyana
University of North Carolina at Greensboro (UNCG)
n_bukhar@uncg.edu

Bulut, Okan
University of Alberta
bulut@ualberta.ca

Burke, Matthew
National Commission on Certification of Physician Assistants
mattb@nccpa.net

Butakor, Paul
University of Alberta
butakor@ualberta.ca

Cai, Li
UCLA
lcai@ucla.edu

Cancado, Luciana
University of Wisconsin-Milwaukee
cancado@uwm.edu

Cao, Chunhua
University of South Florida
chunhuacao@mail.usf.edu

Cappaert, Kevin
University of Wisconsin - Milwaukee
cappaer3@uwm.edu

Carroll, Sarah
Castle Worldwide, Inc.
scarroll@castleworldwide.com

Chang, Yu-Feng
Illinois State Board of Education
chang648@umn.edu

Chang, Shu-Ren
University of Minnesota, Twin Cities
changshuren@huskers.unl.edu

Chang, Yu-Wei
Dept. of Statistics, University of Illinois at Urbana-Champaign
ywchang1225@gmail.com

Chao, Hsiu-Yi
National Chung Cheng University
hsiuyi1118@gmail.com

Cheema, Jehanzeb
University of Illinois at Urbana-Champaign
jrcheema@illinois.edu

Chen, Haiqin
The Ohio State University
chenh@ada.org

Chen, Hui-Fang
City University of Hong Kong
g8932006@gmail.com

Chen, Juan
National Conference of Bar Examiners
jchen@ncbex.org

Contact Information for Individual and Coordinated Sessions First Authors

Chen, Jing
Educational Testing Service
jchen003@ets.org

Chen, Pei-Hua
National Chiao Tung University
peihuamail@gmail.com

Chiang, Yi-Chen
Indiana University Bloomington
chiangy@indiana.edu

Chien, Yuehmei
Pearson
yuehmei.chien@pearson.com

Chiu, Chia-Yi
Rutgers, The State University of New Jersey
chia-yi.chiu@gse.rutgers.edu

Choi, Hye-Jeong
University of Georgia
hjchoi1@uga.edu

Choi, Jinah
University of Iowa
jinah-choi@uiowa.edu

Choi, Seung
CTB McGraw-Hill
seung_choi@ctb.com

Chu, Kwang-Lee
Pearson
kwang-lee.chu@pearson.com

Chubbuck, Kay
Educational Testing Service
kchubbuck@ets.org

Clark, Amy K.
University of Kansas
akclark@ku.edu

Clauser, Jerome
American Board of Internal Medicine
jclauser@abim.org

Cui, Ying
University of Alberta
yc@ualberta.ca

Cui, Zhongmin
ACT Inc
zhongmin.cui@act.org

Culpepper, Steven
University of Illinois at Urbana-Champaign
sculpepp@illinois.edu

Curcin, Milja
Department for Education, Standards and Testing Agency
milja.curcin@education.gsi.gov.uk

Curran, Paul
Kenyon College
curranp@kenyon.edu

Davenport, Jr., Ernest C.
University of Minnesota
lqr6576@umn.edu

Davidson, Anne
Anne H Davidson
anne.davidson@ctb.com

Davis, Laurie L.
Pearson
laurie.davis@pearson.com

De Boeck, Paul
OSU
deboeck.2@osu.edu

De Champlain, Andre
Medical Council of Canada
adechamplain@mcc.ca

Debeer, Dries
University of Leuven
dries.debeer@ppw.kuleuven.be

Delgado-Maldonado, Laura
Instituto Nacional para la Evaluación de la Educación
ldelgado@inee.edu.mx

Demir, Ergül
Ankara University Educational Sciences Faculty
erguldemir@gmail.com

Contact Information for Individual and Coordinated Sessions First Authors

Denbleyker, John
Minnesota Department of Education
lakeway01@yahoo.com

Derickson, Ryan
VHA National Center for Organization Development
rlderickson@gmail.com

Diakow, Ronli
New York University
rd110@nyu.edu

Diao, Hongyu
Umass-Amherst
denisediao@gmail.com

DiBello, Louis V.
University of Illinois at Chicago
ldibello@uic.edu

Ding, Cody
University of Missouri-St. Louis
dingc@umsl.edu

Dogan, Enis
Achieve
edogan@parcconline.org

Donoghue, John
Educational Testing Service
jdonoghue@ets.org

Dunn, Jennifer
Measured Progress
dunn.jennifer@measuredprogress.org

Eckerly, Carol A.
University of Wisconsin-Madison
eckerly@wisc.edu

Embretson, Susan
Georgia Institute of Technology
susan.embretson@psych.gatech.edu

Falk, Carl
University of California, Los Angeles
cffalk@gmail.com

Fan, Fen
University of Massachusetss at Amherst
ffan@educ.umass.edu

Fan, Meichu
ACT
meichu.fan@act.org

Faulkner-Bond, Molly
UMass Amherst
mfaulkne@educ.umass.edu

Feinberg, Richard
National Board of Medical Examiners
rfeinberg@nbme.org

Ferrara, Steve
Pearson Research and Innovation Network
steve.ferrara@pearson.com

Fleishcer, Avi
Illinois Institute of Technology
mfleisch@iit.edu

Flowers, Claudia P.
University of North Carolina, Charlotte
cpflower@uncc.edu

Floyd, Tianna
Georgia State University
tfloyd6@student.gsu.edu

Foelber, Kelly
James Madison University
foelbekj@dukes.jmu.edu

Foley, Brett P.
Alpine Testing Solutions
brett.foley@alpinetesting.com

Foltz, Peter
Pearson
peter.foltz@pearson.com

Forgione, Pascal D.
ETS K-12 Assessment Program
pdforgione@k12center.org

Forte, Ellen
edCount, LLC
eforte@edCount.com

Francis, Xueying
Texas A&M University, College Station
catherine23@neo.tamu.edu

Contact Information for Individual and Coordinated Sessions First Authors

Freeman, Leanne
UW Milwaukee
leannes4@uwm.edu

Frey, Andreas
Friedrich Schiller University Jena
andreas.frey@uni-jena.de

Fujimoto, Ken A.
Loyola University Chicago
kfujimoto@luc.edu

Furgol Castellano, Katherine
Educational Testing Service
KEcastellano@ets.org

G√°ndara, M. Fernanda
University of Massachusetts Amherst
mgandara@educ.umass.edu

Gierl, Mark
University of Alberta
mark.gierl@ualberta.ca

Glaze, Ryan
Pearson
ryan.glaze@pearson.com

Gonzalez, Jorge
Pontificia Universidad Catolica de Chile,
Faculty of Mathematics
jgonzale@mat.puc.cl

Graf, Edith A.
Educational Testing Service
agraf@ets.org

Greive, Elizabeth
NC State University
elgreive@ncsu.edu

Grochowalski, Joseph
Fordham University
jgrochowalsk@fordham.edu

Guo, Lei
Southwest University
happygl1229@gmail.com

Guo, Qi
University of Alberta
qig@ualberta.ca

Guo, Rui
University of Illinois at Urbana-Champaign
ruiguo1@illinois.edu

Habing, Brian
University of South Carolina
habing@stat.sc.edu

Halpin, Peter
NYU
peter.halpin@nyu.edu

Hansen, Mary A.
Robert Morris University
hansen@rmu.edu

Harik, Polina
National Board of Medical Examiners
pharik@nbme.org

Haring, Samuel H.
University of Texas at Austin
samuel.haring@utexas.edu

Harrell, Lauren
University of California - Los Angeles
laurenharrell@ucla.edu

Harris, Deborah
ACT, Inc
deborah.harrtis@act.org

Hauenstein, Clifford E.
Georgia Institute of Technology
cehiv87@gmail.com

He, Qiwei Britt
Educational Testing Service
qhe@ets.org

He, Yong
ACT Inc
yong.he@act.org

Hembry, Ian F.
Amplify Education
ian.hembry@gmail.com

Hendrix, Leslie
University of South Carolina
leslieahendrix@gmail.com

Contact Information for Individual and Coordinated Sessions First Authors

Henson, Robert A.
University of North Carolina at Greensboro
rahenson@uncg.edu

Ho, Andrew
Harvard Graduate School of Education
Andrew_Ho@gse.harvard.edu

Ho, Tsung-Han
ETS
tho@ets.org

Hodge, Kari J.
Baylor University
kari_hodge@baylor.edu

Hollingshead, Lynne
York Region District School Board
lynne.hollingshead@mail.utoronto.ca

Hong, Yuan
American Institutes for Research
yhong@air.org

Hou, Likun
ETS
lhou@ets.org

Hsu, Chia-Ling
The Hong Kong Institute of Educaiotn
jalin518@gmail.com

Huang, Hung-Yu
University of Taipei
hyhuang@go.utaipei.edu.tw

Huff, Kristen
Regents Research Fund
Kristen.Huff@nysed.gov

Huggins-Manley, Anne Corinne
University of Florida
ahuggins@coe.ufl.edu

Huh, Nooree
ACT, Inc
nooree.huh@act.org

HUNG, SU-PIN
National Cheng Kung University
suping0612@gmail.com

Hurtz, Greg
PSI Services LLC
ghurtz@psionline.com

Iaconangelo, Charles
Rutgers, the State University of New Jersey
charles.iaconangelo@gmail.com

Im, Suk Keun
University of Kansas
sukkeun@ku.edu

Janssen, Rianne
KU Leuven
rianne.janssen@ppw.kuleuven.be

Jiang, Yanming
ETS
YXJiang@ets.org

Jiao, Hong
University of Maryland
hjiao@umd.edu

Jin, Kuan-Yu
The Hong Kong Institute of Education
kyjin@ied.edu.hk

Jin, Ying
Association of American Medical Colleges
ying.jin@mtsu.edu

Johnson, Matthew
Teachers College
johnson@tc.edu

Joo, Seang-Hwane
University of South Florida
sjoo@mail.usf.edu

Kaduk, Catherine
University of Illinois at Chicago
ckaduk2@uic.edu

Kang, Chansuk
University of Nebraska-Lincoln
coldstone78@gmail.com

Kang, Hyeon-Ah
University of Illinois - Urbana Champaign
hkang31@illinois.edu

Contact Information for Individual and Coordinated Sessions First Authors

Kang, Yulim
Yonsei University
kangyulim@naver.com

Kaniskan, Burcu
NCBE
burcukaniskan@gmail.com

Kannan, Priya
Educational Testing Service
pkannan@ets.org

Kao, Shu-Chuan
Pearson Vue
shu-chuan.kao@pearson.com

Karadavut, Tugba
UGA
tugba-mat@hotmail.com

Katz, Irvin R.
Educational Testing Service
ikatz@ets.org

Kaya, Yasemin
University of Florida
yaseminkaya@ufl.edu

Keller, Robert
Measured Progress, Inc
commercial@robkeller.com

Kern, Justin L.
University of Illinois at Urbana-Champaign
kern4@illinois.edu

Kettler, Ryan J.
Rutgers, The State University of New Jersey
r.j.kettler@rutgers.edu

Khademi, Abdolvahab
University of Massachusettes
vahab.khademi@gmail.com

Khorramdel, Lale
Educational Testing Service
lkhorramdel@ets.org

Kieftenbeld, Vincent
CTB/McGraw-Hill Education
vincent.kieftenbeld@ctb.com

Kim, Dong-In
CTB/McGraw Hill
dong-in.kim@ctb.com

Kim, Hyung Jin
University of Iowa
hyungjin-kim@uiowa.edu

Kim, Ja Young
ACT, Inc.
jayoung.kim@act.org

Kim, Jungnam
NBCE
jungnam95@hotmail.com

Kim, Han Yi
Measured Progress
Kim.HanYi@measuredprogress.org

Kim, Kyung Yong
University of Iowa
kyungyong-kim@uiowa.edu

Kim, Min Sung
University of Kansas
mskim@ku.edu

King, David
Georgia Tech
david.randall.king@gatech.edu

King, John
USED
John.King@ed.gov

Kleper, Dvir
National Institute for Testing and Evaluation
dvir@nite.org.il

Koenig, Judith A.
National Academy of Science/
National Research Council
jkoenig@nas.edu

Koo, Jin
American Nurses Credentialing Center
koo.jin@yahoo.com

Kretschmann, Julia
University of Potsdam, Germany
julia.kretschmann@uni-potsdam.de

Contact Information for Individual and Coordinated Sessions First Authors

Kunze, Katie L.
Arizona State University
katie.kunze@asu.edu

Kyllonen, Patrick
ETS
pkyllonen@ets.org

LaFond, Lee
Measured Progress
lafond.lee@measuredprogress.org

Lai, Hollis
University of Alberta
hollis.lai@ualberta.ca

LAN, MING-CHIH
University of Washington
mclan@uw.edu

Lane, Suzanne
University of Pittsburgh
sl@pitt.edu

Latifi, Syed Muhammad Fahad
University of Alberta
fahad.latifi@ualberta.ca

Lee, Chansoon
clee284@wisc.edu

Lee, Juyeon
Yonsei University
k3jle69@naver.com

Lee, Philseok
University of South Florida
philseok@mail.usf.edu

Lee, Won-Chan
University of Iowa
won-chan-lee@uiowa.edu

Lee, Woo-yeol
Vanderbilt University
woo-yeol.lee@vanderbilt.edu

Lee, Yoonsun
Seoul women's University
ylee@swu.ac.kr

Leighton, Jacqueline
University of Alberta
jacqueline.leighton@ualberta.ca

Leventhal, Brian
University Of Pittsburgh
brl38@pitt.edu

Li, Dongmei
ACT Inc
dongmei.li@act.org

Li, Feiming
University of North Texas Health Science Center
feiming.li@unthsc.edu

Li, Haiying
University of Wisconsin
haiyinglit@gmail.com

Li, ChengHsien
Michigan State University
lichengh@msu.edu

Li, Zhen
University of California, Los Angeles
lizhenjuly@ucla.edu

Li, Xiaomin
The Hong Kong Institute of Education
nickylxm@yahoo.com.hk

Li, Tianli
ACT Inc.
tianli.li@act.org

Li, Tongyun
Educational Testing Service
tli002@ets.org

Li, Xiaoran
University of Connecticut
xiaoran.li@uconn.edu

Li, Xin
ACT, Inc.
xin.li@act.org

Li, Zhushan
Boston College
zhushan.li@gmail.com

Contact Information for Individual and Coordinated Sessions First Authors

Liao, Dandan
University of Maryland
echommm@gmail.com

Liaw, Yuan-Ling
University of Washington
linda08@uw.edu

Lim, Euijin
The University of Iowa
euijin-lim@uiowa.edu

Lin, Haiyan
Act, Inc.
haiyan.lin@act.org

Lin, Johnny
Educational Testing Service
jlin@ets.org

Lin, Zhe
Beijing Normal University
lz_psy@163.com

Lin, Peng
ETS
plin@ets.org

ling, guangming
Educational Testing Service
gling@ets.org

List, Marit Kristine
IPN - Leibniz Institute of Science and Mathematics Education
list@ipn.uni-kiel.de

Liu, Chunyan
ACT, Inc.
chunyan.liu@act.org

Liu, Yang
The University of North Carolina at Chapel Hill
liuy0811@live.unc.edu

Liu, Ou Lydia
ETS
lliu@ets.org

Liu, Yixing
Arizona State University
yixing.liu@asu.edu

Lockwood, John
Educational Testing Service
jrlockwood@ets.org

Longabach, Tanya
University of Kansas
tlongabach@ku.edu

Loughran, Jessica
University of Kansas
jtl@ku.edu

Lu, Yang
ACT, inc.
yang.lu@act.org

Lu, Ying
Educational Testing Service
ylu@ets.org

Luhanga, Ulemu
Queen's University
ulemuluhanga@gmail.com

Luna Bazaldua, Diego A.
Teachers College, Columbia University
dal2159@tc.columbia.edu

Luo, Xin
Michigan State University
luoxin1@msu.edu

Ma, Wenchao
Graduate School of Education
wenchao.ma@rutgers.edu

Mao, Liyang
Educational Testing Service
maoliyan@msu.edu

Margolis, Melissa
National Board of Medical Examiners
margolis@nbme.org

Marino, Katherine
Pennsylvania State University
katemarino2@gmail.com

Marland, Joshua
University of Massachusetts Amherst
joshua.marland@gmail.com

Contact Information for Individual and Coordinated Sessions First Authors

Martin, Michelle
Educational Testing Service
mmartin001@ets.org

Martinez Alpizar, David
CSUN
damartinezalpizar.43@gmail.com

Matta, Tyler H.
University of Oregon
tmatta@uoregon.edu

McClarty, Katie
Pearson
katie.mcclarty@pearson.com

McCoy, Thomas P.
UNC Greensboro
tpmccoy@uncg.edu

McLeod, Jeffrey T.
Pearson
jeff.mcleod@pearson.com

Meyer, Patrick
University of Virginia
jpm4qs@virginia.edu

Michaelides, Michalis
University of Cyprus
michalim@ucy.ac.cy

Michel, Rochelle
ETS
rmichel@ets.org

Minchen, Nathan
Rutgers, The State University of New Jersey
nathan.minchen@rutgers.edu

Mittelhaeuser, Marie-Anne
Cito
Marie-Anne.Mittelhaeuser@cito.nl

Mo, Ya
Michigan State University
moya@msu.edu

Monfils, Lora
ETS
lmonfils@ets.org

Monroe, Scott
UCLA
scott.monroe@ucla.edu

Moore, Christopher
Minneapolis Public Schools
moor0554@umn.edu

Mroch, Andrew A.
ACT
andrew.mroch@act.org

Muntean, William
Pearson
williamjmuntean@gmail.com

Naumann, Johannes
Goethe-University
j.naumann@em.uni-frankfurt.de

Naumann, Alexander
German Institute for International
Educational Research
naumanna@dipf.de

Naumenko, Oksana
The University of North Carolina at Greensboro
o_naumen@uncg.edu

Neidorf, Teresa
American Institutes for Research
tneidorf@air.org

Nese, Joseph
University of Oregon
jnese@uoregon.edu

Nicewander, Alan
Pacific Metrics
alan.nicewander@gmail.com

Nickodem, Kyle
University of Minnesota
nicko013@umn.edu

O'Leary, Timothy M.
University of Melbourne
t.oleary@student.unimelb.edu.au

Oh, Hyeon-Joo
Educational Testing Service
hoh@ets.org

Contact Information for Individual and Coordinated Sessions First Authors

Olivera Aguilar, Margarita
Educational Testing Service
margarita.olag@gmail.com

Oliveri, Maria Elena
Educational Testing Service
moliveri@ets.org

Olsen, James
Renaissance Learning
jamesbolsen@hotmail.com

Olson, John F.
Olson Educational Measurement & Assessment Services
jmclkolson@yahoo.com

Oluwalana, Olasumbo
Rutgers University
oluwalan@scarletmail.rutgers.edu

Oranje, Andreas
ETS
aoranje@ets.org

Padilla, Jose-Luis
University of Granada
jpadilla@ugr.es

Pak, Seohong
University of Iowa
seohong-pak@uiowa.edu

Park, Jiyoon
Federation of State Boards of Physical Therapy
jpark@fsbpt.org

Park, Jung Yeon
Teachers College, Columbia University
jyp2111@tc.columbia.edu

Park, Hye-Sook
Honam University
parkhyes@honam.ac.kr

Park, Yoon Soo
University of Illinois at Chicago
yspark2@uic.edu

Patelis, Thanos
Center for Assessment
tpatelis@nciea.org

Patterson, Brian
Pearson Education, Inc.
brian.f.patterson@gmail.com

Peng, Luyao
University of California Riverside
lpeng002@ucr.edu

Perie, Marianne
Center for Educational Testing and Evaluation
mperie@ku.edu

Peterson, Jaime
University of Iowa
jaime-peterson@uiowa.edu

Phillips, S E
sepsearch@aol.com

Poggio, John
University of Kansas
jpoggio@ku.edu

Pokropek, Artur
IFiS PAN
artur.pokropek@gmail.com

Popham, William J.
University of California Los Angeles

Puhan, Gautam
ETS
gpuhan@ets.org

Qian, Jiahe
Educational Testing Service
jqian@ets.org

QIU, Xue-Lan
Hong Kong Institute of Education
xlqiu@ied.edu.hk

Ramineni, Chaitanya
Educational Testing Services
cramineni@ets.org

Randall, Jennifer
University of Massachusetts
jrandall@educ.umass.edu

Contact Information for Individual and Coordinated Sessions First Authors

Rankin, Angelica
University of Iowa
Angelica-Rankin@uiowa.edu

Raymond, Mark
National Board of Medical Examiners
mraymond@nbme.org

Rickels, Heather A.
University of Iowa
heather-rickels@uiowa.edu

Rijmen, Frank
CTB
frank.rijmen@ctb.com

Rios, Joseph
University of Massachusetts, Amherst
jarios@educ.umass.edu

Roberts, James S.
Georgia Institute of Technology
james.roberts@psych.gatech.edu

Robin, Frederic
ETS
frobin@ets.org

Rodriguez, Michael
University of Minnesota
mcrdz@umn.edu

Rogers, H Jane
University of Connecticut
jane.rogers@uconn.edu

Roohr, Katrina C.
Educational Testing Service
KRoohr@ets.org

Rutkowski, David
Indiana University
davidrutkowski@gmail.com

Rutkowski, Leslie
Indiana University
lrutkows@indiana.edu

Sachse, Karoline A.
Institute for Educational Quality Improvement,
Humboldt-University of Berlin
sachseka@hu-berlin.de

Sahin, Alper
Cankaya University
asahin@cankaya.edu.tr

Saiar, Amin
PSI Services LLC
amin@psionline.com

Sano, Makoto
Prometric Inc.
makoto.sano@prometric.com

Sato, Edynn
Pearson
edynn.sato@pearson.com

Sen, Sedat
The University of Georgia
sedatsen06@gmail.com

Seo, Minhee
Korea Institute for Curriculum & Evaluation
minicap@gmail.com

Seo, Dong Gi
National Registry of Emergency
Medical Technicians
wmotive@gmail.com

Sgammato, Adrienne
ETS
asgammato09@gmail.com

Sha, Shuying
University of North Carolina at Greensboro
s_sha@uncg.edu

Sharma, Anu
University of Kansas
anusharma@ku.edu

Shear, Benjamin R.
Stanford University
benjamin.shear@gmail.com

Sherlock, Phillip R.
University of South Carolina
sherlocp@mailbox.sc.edu

Shim, Hi Shin
Georgia Institute of Technology
hishin@gatech.edu

Contact Information for Individual and Coordinated Sessions First Authors

Shin, AhYoung
University of Iowa
ahyoung-shin@uiowa.edu

Shin, MinJeong
American Institutes For Research
mshin@air.org

Shin, Nami
UCLA
nami0623@gmail.com

Shulruf, Boaz
University of New South Wales
b.shulruf@unsw.edu.au

Sikali, Emmanuel
National Center for Education Statistics
Emmanuel.Sikali@ed.gov

Sinharay, Sandip
CTB/McGraw-Hill
sandip_sinharay@ctb.com

Skorupski, William P.
University of Kansas
wps@ku.edu

Smith, Weldon Z.
University of Nebraska-Lincoln
weldon@huskers.unl.edu

Sorrel, Miguel
Universidad Aut√≥noma de Madrid
sorrel.mig@gmail.com

Soto, Amanda
National Board of Medical Examiners
asoto@nbme.org

Stevens, Joseph J.
University of Oregon
stevensj@uoregon.edu

Stevenson Jr, Zollie
Howard University
zstevenson@aol.com

Stone, Elizabeth
ETS
estone@ets.org

Straat, Hendrik
Cito
hendrik.straat@cito.nl

Strauts, Erin
erin.strauts@gmail.com

Sukin, Tia M.
Pacific Metrics
tsukin@pacificmetrics.com

Sussman, Joshua M.
UC Berkeley
jsussman@berkeley.edu

Swinburne Romine, Russell
University of Kansas
swin0030@ku.edu

Tan, Xuan (Adele)
ETS
atan@ets.org

Tang, Shuwen
UW-Milwaukee
tangsw.1106@gmail.com

Tang, Wei
University of Alberta
wtang3@ualberta.ca

Tannenbaum, Richard
Educational Testing Service
rtannenbaum@ets.org

Tao, Wei
ACT, Inc.
taowei3@gmail.com

Terzi, Ragip
Rutgers, The State University of New Jersey
terziragip@gmail.com

Thissen, David
University of North Carolina
dthissen@email.unc.edu

Tian, Wei
Beijing Normal University
tianwei65396@163.com

Contact Information for Individual and Coordinated Sessions First Authors

Topczewski, Anna
Pearson
anna.topczewski@pearson.com

Trierweiler, Tammy
Prometric
tjtrier@gmail.com

Tzou, Hueying
National University of Tainan
tzou@mail.nutn.edu.tw

van der Linden, Wim J.
CTB/McGraw-Hill
wim_vanderlinden@ctb.com

Van Nijlen, Daniel
KULeuven BE0419.052.173
daniel.vannijlen@ppw.kuleuven.be

van Rijn, Peter
ETS Global
pvanrijn@etsglobal.org

Vispoel, Walter P.
University of Iowa
walter-vispoel@uiowa.edu

von Davier, Alina
ETS
avondavier@ets.org

Waldman, Marcus
Harvard Grad. School. of Ed.
mrw484@mail.harvard.edu

Walker, Cindy M.
University of Wisconsin-Milwaukee
cmwalker@uwm.edu

Wan, Lei
Pearson
lei.wan@pearson.com

Wang, Yang
Education Analytics
awangyang@gmail.com

Wang, Changjiang
Pearson
Changjiang.Wang@Pearson.com

Wang, Hongling
ACT, Inc.
hongling.wang@act.org

Wang, Zhen
Educational Testing Service
jwang@ets.org

Wang, Keyin
Michigan State University
keyinw0323@gmail.com

Wang, Lu
The University of Iowa
lu-wang-3@uiowa.edu

Wang, Shichao
The University of Iowa
shichao-wang@uiowa.edu

Wang, Wenhao
University of Kansas
wwh8623@gmail.com

Weeks, Jonathan
Educational Testing Service
jweeks@ets.org

Wei, Hua
Pearson
hua.wei@pearson.com

Wei, Youhua
Educational Testing Service
ywei@ets.org

Weirich, Sebastian
Institute for Educational Quality Improvement
sebastian.weirich@iqb.hu-berlin.de

Weiss, David J.
University of Minnesota
djweiss@umn.edu

Wells-Moreaux, Sheila
University of Kansas
sheilawellsmoreaux@ku.edu

Westphal, Andrea
Universität Potsdam
andrea.westphal@uni-potsdam.de

Contact Information for Individual and Coordinated Sessions First Authors

Widiatmo, Heru
ACT, Inc.
heru.widiatmo@act.org

Wiley, Andrew
Alpine Testing Solutions
andrew.wiley@alpinetesting.com

Williams, Elizabeth
Georgia Institute of Technology
ewilliams62@gatech.edu

Williams, Frank
Educational Testing Service
fwilliams001@ets.org

Willse, John T.
University of North Carolina at Greensboro
jtwillse@uncg.edu

Wise, Lauress
HumRRO
lwise@HumRRO.org

Wise, Steven
Northwest Evaluation Association
steve.wise@nwea.org

Wolfe, Edward
Pearson
ed.wolfe@pearson.com

Wolkowitz, Amanda
Alpine Testing
amanda.wolkowitz@alpinetesting.com

Wood, Scott W.
Pacific Metrics Corporation
swood@pacificmetrics.com

Wyse, Adam
American Registry of Radiologic Technologists
adam.wyse@arrt.org

Xie, Aolin
Prometric
olymxie@gmail.com

Xing, Kuan
University of Illinois at Chicago
kuanxing83@gmail.com

Xu, Wei
University of Florida
x.wei1007@gmail.com

Xu, Jing-Ru
Michigan State Univ
xujingru@msu.edu

Yan, Duanli
ETS
dyan@ets.org

Yan, Ning
Independent Consultant
Ning.Yan@pearson.com

Yang, Fan
Pearson/The University of Iowa
fan-yang-3@uiowa.edu

Yang, Ping
The University of Iowa
pyq3b@mail.missouri.edu

Yang, Sujin
Yonsei University
renewsjlife@gmail.com

Yang, Chien-Lin
University of Missouri-Columbia
yangc@ada.org

Yang, Zhiming
Educational Records Bureau
yzm506jx@yahoo.com

Yao, Lihua
Defense Manpower Data Center
Lihua.Yao.civ@mail.mil

Yavuz, Guler
Hacettepe University
rkh@educ.umass.edu

Ye, Sam
University of Illinois - Urbana Champaign
sye3@illinois.edu

Yel, Nedim
Arizona State University
nedimyel@gmail.com

Contact Information for Individual and Coordinated Sessions First Authors

Yen, Shu Jing
Center for Applied Linguistics
syen@cal.org

Yilmaz, Mustafa
The University of Kansas
myilmaz@ku.edu

Yoon, Jiyoung
Seoul Women,Äôs University
ellie5900@naver.com

Yoon, Su-Youn
Educational Testing Service
syoon@ets.org

Yu, Hsiu-Ting
McGill University
hsiutingyu@gmail.com

Zahner, Doris
CAE
dzahner@cae.org

Zapata-Rivera, Diego
Educational Testing Service
dzapata@ets.org

Zenisky, April
University of Massachusetts Amherst
azenisky@educ.umass.edu

Zhang, Jin
ACT
jin.zhang@act.org

Zhang, Jinming
University of Illinois at Urbana-Champaign
jmzhang@illinois.edu

Zhang, Liru
Deleware State Department of Education
liru.zhang@doe.k12.de.us

Zhang, Mengyao
University of Iowa
mengyao-zhang@uiowa.edu

Zhang, Mo
Educational Testing Service
mzhang@ets.org

Zhang, Xiuyuan
The College Board
xzhang@collegeboard.org

Zhang, Jiahui
Michigan State University
zhang321@msu.edu

Zhang, Mingcai
Michigan State University
zhangmc@msu.edu

Zheng, Xiaying
EDMS, University of Maryland, College Park
xyzheng86@gmail.com

Zhou, Xuechun
Pearson
xuechun.zhou@pearson.com

Ziker, Cindy
SRI International
Cindy.Ziker@sri.com

Zopluoglu, Cengiz
University of Miami
c.zopluoglu@miami.edu

Zu, Jiyun
Educational Testing Service
jzu@ets.org

Zwick, Rebecca
Educational Testing Service
rzwick@cox.net

NCME 2015 • Schedule-At-A-Glance

Time	Room	Type	ID	Title
Wednesday, April 15, 2015				
8:00 AM-12:00 PM	Exchange (11th Floor)	TS	AA	A Practitioner's Guide to Growth Models
8:00 AM-12:00 PM	Seville Ballroom West (Lobby Level)	TS	BB	An Introduction to Equating in R
8:00 AM-12:00 PM	Valencia (Lobby Level)	TS	CC	Using Visual Displays to Inform Assessment Development and Validation
8:00 AM-5:00 PM	Empire Ballroom (7th Floor)	TS	DD	Leveraging Open Source Software and Tools for Statistics/Measurement Research
8:00 AM-5:00 PM	Renaissance Ballroom (5th Floor)	TS	EE	flexMIRT®: Flexible Multilevel Multidimensional Item Analysis and Test Scoring
8:00 AM-5:00 PM	Seville Ballroom East (Lobby Level)	TS	FF	An Introduction to Diagnostic Classification Modeling
8:00 AM-5:00 PM	Toledo (5th Floor)	TS	GG	Optimal Test Design
1:00 PM-5:00 PM	Exchange (11th Floor)	TS	HH	An Overview of Operational Psychometric Work in Real World
1:00 PM-5:00 PM	Seville Ballroom West (Lobby Level)	TS	II	A Graphical and Nonlinear Mixed Model Approach to IRT with the R Package Flirt
1:00 PM-5:00 PM	Valencia (Lobby Level)	TS	JJ	Cognitive Lab Techniques: An Overview, Framework, and Some Practice
Thursday, April 16, 2015				
8:00 AM-12:00 PM	Exchange (11th Floor)	TS	KK	Fundamentals of Item Response Theory and Computerized Adaptive Testing
8:00 AM-12:00 PM	Seville Ballroom East (Lobby Level)	TS	LL	Item Response Theory With jMetrik and Psychometric Programming With Java
8:00 AM-12:00 PM	Seville Ballroom West (Lobby Level)	TS	MM	Landing Your Dream Job for Graduate Students
8:00 AM-5:00 PM	Empire Ballroom (7th Floor)	TS	OO	Multidimensional Item Response Theory: Theory and Applications and Software
8:00 AM-5:00 PM	Renaissance Ballroom (5th Floor)	TS	PP	Generalizability Theory and Applications
8:00 AM-5:00 PM	Toledo (5th Floor)	TS	QQ	Bayesian Networks in Educational Assessment
1:00 PM-5:00 PM	Exchange (11th Floor)	TS	RR	Advances in Measuring 21st Century Skills: Constructs, Development, and Scoring

CS=Coordinated Session • IS=Invited Session
TS=Training Session • PS=Paper Session
EB=Electronic Board Session

1:00 PM-5:00 PM	Seville Ballroom East (Lobby Level)	TS	SS	Using IRT for Standard Setting in Performance Based Assessments
1:00 PM-5:00 PM	Seville Ballroom West (Lobby Level)	TS	TT	An Introduction to Using R for Quantitative Methods
1:00 PM-5:00 PM	Valencia (Lobby Level)	TS	UU	Understanding Automated Scoring: Theory and Practice
4:00 PM-7:00 PM	Cordova Room (5th Floor)			NCME Board of Directors Meeting
Friday, April 17, 2015				
8:15 AM-10:15 AM	Empire Ballroom (7th Floor)	CS	A1	Use of Evidence-Based Standard Setting in PARCC Assessments
8:15 AM-10:15 AM	Exchange (11th Floor)	PS	A2	DIF: Bayesian and Mixed
8:15 AM-10:15 AM	Grand Ballroom (7th Floor)	CS	A3	Various Efforts to Evaluate the Quality of Assessment Programs
8:15 AM-10:15 AM	King Arthur (3rd Floor)	CS	A4	Test Score Integrity in the Age of Common-Core Assessments
8:15 AM-10:15 AM	Renaissance Ballroom (5th Floor)	IS	A5	NCME-NATD Symposium: Implementing the Common Core Assessments at the District and School Levels: Voices from the Field - Overcoming Challenges, Making it Work
8:15 AM-10:15 AM	Seville Ballroom East (Lobby Level)	CS	A6	Overview: Theories of Action for Performance Assessment in Large Scale Testing Programs
8:15 AM-10:15 AM	Seville Ballroom West (Lobby Level)	PS	A7	Item Development
8:15 AM-10:15 AM	Toledo (5th Floor)	CS	A8	Test Batteries Under Sequential Designs: A Technology-Enhanced Examination
8:15 AM-10:15 AM	Valencia (Lobby Level)	PS	A9	Linking in General
10:35 AM-12:05 PM	Grand Ballroom (7th Floor)	IS	B1	The Role of the Measurement Profession in the Renewal of ESEA and Other Federal Education Initiatives
12:25 PM-1:25 PM	Camelot (3rd Floor)	EB	C1	
12:25 PM-1:55 PM	King Arthur (3rd Floor)	IS	C2	Spencer Foundation: From Funded to Unfunded: What Makes the Difference
12:25 PM-1:55 PM	Renaissance Ballroom (5th Floor)	CS	C3	Measuring Students' Proficiency on the Next Generation Science Standards
12:25 PM-1:55 PM	Toledo (5th Floor)	CS	C4	Automated Scoring, Item Generation and Mixed-Format Adaptive Testing

CS=Coordinated Session • IS=Invited Session
TS=Training Session • PS=Paper Session
EB=Electronic Board Session

12:25 PM-1:55 PM	Seville Ballroom East (Lobby Level)	CS	C5	Methodological Developments in International Large-Scale Assessments
12:25 PM-1:55 PM	Exchange (11th Floor)	IS	C6	Handbook of Test Development (2nd Ed): Major Advances and Implications for Test Developers
12:25 PM-1:55 PM	Empire Ballroom (7th Floor)	IS	C7	Model Fit and Scoring Invariance Across Multiple Populations
12:25 PM-1:55 PM	Grand Ballroom (7th Floor)	IS	C8	Quality Focus: Experiences From a Number of Assessment Programs
12:25 PM-1:55 PM	Valencia (Lobby Level)		C9	Peer Review of Peer Review
12:25 PM-1:55 PM	St. Clair (Upper 5th Floor)	PS	C10	Considerations for Measuring Item Difficulty
2:15 PM-3:45 PM	Camelot	EB	D1	GSIC Poster Session
2:15 PM-3:45 PM	Empire Ballroom (7th Floor)	CS	D2	Applications of Model-Based Rater Monitoring Procedures
2:15 PM-3:45 PM	Exchange (11th Floor)	CS	D3	Assessment for Innovative Learning Technology: Modeling Sources of Dependence
2:15 PM-3:45 PM	Grand Ballroom (7th Floor)	IS	D4	Advances in Test Score Reporting
2:15 PM-3:45 PM	King Arthur (3rd Floor)	CS	D5	Improving Test Security for State Assessment Programs: Lessons Learned
2:15 PM-3:45 PM	Renaissance Ballroom (5th Floor)	CS	D6	Two Approaches to Game Based Assessments: Mods and Originals
2:15 PM-3:45 PM	Seville Ballroom East (Lobby Level)	CS	D7	Methods for Comparing NAEP Frameworks to Other Assessments and Standards
2:15 PM-3:45 PM	Seville Ballroom West (Lobby Level)	CS	D8	Pseudo Equivalent Groups Linking in Large Scale Assessment
2:15 PM-3:45 PM	Toledo (5th Floor)	CS	D9	Reliability, Internal Consistency, and Unidimensionality Related but Distinct Concepts
2:15 PM-3:45 PM	Valencia (Lobby Level)	CS	D10	Beyond Scoring: Alternative Use of Automated Systems for Language Assessments
4:05 PM-6:05 PM	King Arthur (3rd Floor)	IS	E1	Contemporary Problems in Educational Measurement (Satirical Session)
6:30 PM-8:00 PM	Seville Ballroom (Lobby Level)			NCME and AERA Division D Joint Reception

CS=Coordinated Session • IS=Invited Session
TS=Training Session • PS=Paper Session
EB=Electronic Board Session

Saturday, April 18, 2015				
8:00 AM-9:00 AM	Grand Ballroom Salon II, Chicago Marriott Downtown Hotel			NCME Business Meeting and Breakfast
9:00 AM-9:40 AM	Grand Ballroom Salon II, Chicago Marriott Downtown Hotel		IS	NCME Presidential Address
10:35 AM-12:05 PM	Empire Ballroom (7th Floor)	CS	F1	Using Ordered Probit Models to Reconstruct Coarsened Test-Score Distributions
10:35 AM-12:05 PM	Exchange (11th Floor)	CS	F2	Recent Advances and Comparisons of Teacher Effectiveness Models
10:35 AM-12:05 PM	Grand Ballroom (7th Floor)	CS	F3	A Potentially Potent Assessment-Literacy Initiative: Reactions Sought
10:35 AM-12:05 PM	King Arthur (3rd Floor)	IS	F4	NCME Career Award Presentation: Item Response Theory, Serendipity, and Bad Questions
10:35 AM-12:05 PM	Renaissance Ballroom (5th Floor)	PS	F5	Setting Cut Scores
10:35 AM-12:05 PM	Seville Ballroom West (Lobby Level)	CS	F6	Psychometric Considerations for the Next Generation of Performance Assessment
10:35 AM-12:05 PM	Toledo (5th Floor)	PS	F7	Equating Approaches/Methods
10:35 AM-12:05 PM	Valencia (Lobby Level)	PS	F8	CAT for Diagnostic Purposes
12:25 PM-1:25 PM	Camelot (3rd Floor)	EB	G1	
12:25 PM-1:55 PM	Empire Ballroom (7th Floor)	PS	G2	Assessing Diverse Learners
12:25 PM-1:55 PM	Exchange (11th Floor)	PS	G3	Automated Scoring and Text Generation
12:25 PM-1:55 PM	Grand Ballroom (7th Floor)	CS	G4	Ensuring Content Validity and Alignment of Computer Adaptive Reading Assessments
12:25 PM-1:55 PM	King Arthur (3rd Floor)	PS	G5	Technical Investigation of SGPs/VAMs for Teacher Evaluation
12:25 PM-1:55 PM	Renaissance Ballroom (5th Floor)	PS	G6	Performance Level Descriptors
12:25 PM-1:55 PM	Seville Ballroom East (Lobby Level)	PS	G7	Irregularities in Operational Testing
12:25 PM-1:55 PM	Seville Ballroom West (Lobby Level)	PS	G8	Automated Scoring

CS=Coordinated Session • IS=Invited Session
TS=Training Session • PS=Paper Session
EB=Electronic Board Session

12:25 PM-1:55 PM	Toledo (5th Floor)	PS	G9	Equating Methods
12:25 PM-1:55 PM	Valencia (Lobby Level)	PS	G10	Methods for Investigating Threats to Validity
2:15 PM-3:45 PM	Camelot (3rd Floor)	EB	H1	GSIC Poster Session
2:15 PM-3:45 PM	Empire Ballroom (7th Floor)	IS	H2	Measurement and Implementation Challenges in Early Childhood Assessment
2:15 PM-3:45 PM	Exchange (11th Floor)	CS	H3	Issues in Human Scoring of Constructed-Response Items
2:15 PM-3:45 PM	Grand Ballroom (7th Floor)	CS	H4	Evaluating and Improving Methods for Student Growth Percentile Estimation
2:15 PM-3:45 PM	Renaissance Ballroom (5th Floor)	IS	H5	The Importance of Instructional Sensitivity: A Colloquy Among Combatants
2:15 PM-3:45 PM	King Arthur (3rd Floor)	CS	H6	Smarter Balanced Automated Scoring Research: Results and Insights
2:15 PM-3:45 PM	Seville Ballroom East (Lobby Level)	CS	H7	Third Grade Reading Proficiency: Two Large-Scale Longitudinal Studies
2:15 PM-3:45 PM	Seville Ballroom West (Lobby Level)	CS	H8	Feasibility of Various Cut Score Moderation Methods
2:15 PM-3:45 PM	Toledo (5th Floor)	CS	H9	Research and Development on Assessment and Accountability for Special Education
2:15 PM-3:45 PM	Valencia (Lobby Level)	PS	H10	Smoothing in Equating
4:05 PM-5:05 PM	Camelot (3rd Floor)	EB	I1	
4:05 PM-6:05 PM	St. Clair (Upper 5th Floor)	CS	I2	Evaluating Scoring Issues for Innovative and Technology Enhanced Items
4:05 PM-6:05 PM	Exchange (11th Floor)	CS	I3	Psychometrics in a Learning Maps Environment
4:05 PM-6:05 PM	Seville Ballroom East (Lobby Level)	IS	I4	A Dialogue for Addressing Measurement and Data Gaps in Education
4:05 PM-6:05 PM	King Arthur (3rd Floor)	CS	I5	Surf and Turf Summative Assessment: States Combining Efficiencies With Customization
4:05 PM-6:05 PM	Renaissance Ballroom (5th Floor)	IS	I6	Standard Setting in the Common Core World: PARCC and SBAC Experiences
4:05 PM-6:05 PM	Adler (2nd Floor)	PS	I7	Person Fit and Aberrant Responses
4:05 PM-6:05 PM	Seville Ballroom West (Lobby Level)	PS	I8	DIF: Sample Size, Effect Size, Power
4:05 PM-6:05 PM	Toledo (5th Floor)	PS	I9	Extraneous Factors Affecting Test Behaviors
4:05 PM-6:05 PM	Valencia (Lobby Level)	PS	I10	Investigations in Examinee Guessing and Response Time

CS=Coordinated Session • IS=Invited Session
TS=Training Session • PS=Paper Session
EB=Electronic Board Session

Sunday, April 19, 2015				
5:45 AM-7:00 AM	Meet at the InterContinental Hotel Lobby			NCME Fitness Run / Walk
6:30 AM-7:30 AM	Grand Ballroom Balcony (8th Floor)			Yoga
8:15 AM-10:15 AM	Empire Ballroom (7th Floor)	PS	J1	Improving Proficiency Estimation
8:15 AM-10:15 AM	Exchange (11th Floor)	PS	J2	Performance Scoring Using Raters and Constructed Responses
8:15 AM-10:15 AM	Grand Ballroom (7th Floor)	CS	J3	Innovative Perspectives on Common-Core Tests: PARCC & SBAC Compare Notes
8:15 AM-10:15 AM	King Arthur (3rd Floor)	PS	J4	Detecting Bias Across Special Populations
8:15 AM-10:15 AM	Renaissance Ballroom (5th Floor)	CS	J5	Gathering and Evaluating Validity Evidence Based on Response Processes
8:15 AM-10:15 AM	Seville Ballroom East (Lobby Level)	IS	J6	Exploring the Implications of the "Fairness" Chapter of the 2014 Standards for Educational and Psychological Testing
8:15 AM-10:15 AM	Seville Ballroom West (Lobby Level)	PS	J7	Multidimensional Item Response Theory
8:15 AM-10:15 AM	Toledo (5th Floor)	CS	J8	Delivering the National Assessment on Tablet: Psychometric Challenges and Opportunities
8:15 AM-10:15 AM	Valencia (Lobby Level)	IS	J9	Awards Session
10:35 AM-12:05 PM	Empire Ballroom (7th Floor)	CS	K1	Multiple Facets of an Assessment With Collaborative Problem Solving Tasks
10:35 AM-12:05 PM	Exchange (11th Floor)	CS	K2	Designing Next-Generation Assessments of Student Learning Outcomes in Higher Education
10:35 AM-12:05 PM	King Arthur (3rd Floor)	CS	K3	Constructing a Vertical Scale Under Linked Scaling Tests Design
10:35 AM-12:05 PM	Renaissance Ballroom (5th Floor)	CS	K4	Do Interruptions During Online Testing Impact the Examinee Scores?
10:35 AM-12:05 PM	Seville Ballroom East (Lobby Level)	CS	K5	Current Issues in Test Assembly
10:35 AM-12:05 PM	Seville Ballroom West (Lobby Level)	CS	K6	Toward More Robust Automated Essay Scoring Models
10:35 AM-12:05 PM	Toledo (5th Floor)	CS	K7	Detection and Solutions of Aberrant Performances of Automated Scoring Systems
10:35 AM-12:05 PM	Valencia (Lobby Level)	PS	K8	Using Validity Evidence in Diverse Settings

CS=Coordinated Session • IS=Invited Session
TS=Training Session • PS=Paper Session
EB=Electronic Board Session

12:25 PM-1:25 PM	Camelot (3rd Floor)	EB	L1	
12:25 PM-1:55 PM	Empire Ballroom (7th Floor)	PS	L2	Subscore Reporting
12:25 PM-1:55 PM	Exchange (11th Floor)	PS	L3	Innovations in Teacher Evaluation
12:25 PM-1:55 PM	Grand Ballroom (7th Floor)	CS	L4	Using Principled Assessment Frameworks to Guide Test Development and Validation
12:25 PM-1:55 PM	King Arthur (3rd Floor)	PS	L5	Innovations in Operational Testing
12:25 PM-1:55 PM	Renaissance Ballroom (5th Floor)	PS	L6	DIF With Special Item Types
12:25 PM-1:55 PM	Seville Ballroom West (Lobby Level)	PS	L7	Comparing Equating Methods
12:25 PM-1:55 PM	Toledo (5th Floor)	PS	L8	Linking and Vertical Scaling
12:25 PM-1:55 PM	Valencia (Lobby Level)	PS	L9	Mixture IRT Models
2:15 PM-3:45 PM	Empire Ballroom (7th Floor)	CS	M1	Psychometric Considerations for PARCC Assessments: Research From the Field Test
2:15 PM-3:45 PM	Exchange (11th Floor)	CS	M2	Theory-Based Item Generation for Mathematics Assessment and Instruction
2:15 PM-3:45 PM	Grand Ballroom (7th Floor)	IS	M3	Debate: Equal Interval Scales in Educational Testing: Attainable Goal or Myth?
2:15 PM-3:45 PM	King Arthur (3rd Floor)	PS	M4	Subscore Recovery
2:15 PM-3:45 PM	Renaissance Ballroom (5th Floor)	PS	M5	Reliability Related New Models
2:15 PM-3:45 PM	Seville Ballroom East (Lobby Level)	CS	M6	Psychometric Considerations in Linking to Survey Assessments
2:15 PM-3:45 PM	Seville Ballroom West (Lobby Level)	PS	M7	Score Reporting
2:15 PM-3:45 PM	Toledo (5th Floor)	CS	M8	Thinking About Validity in Measuring Teacher and School Effectiveness
2:15 PM-3:45 PM	Valencia (Lobby Level)	CS	M9	Applications and Advances of Multidimensional IRT With Stochastic Approximation Methods
4:00 PM-7:00 PM	Burnham Room (8th Floor)			NCME Board of Directors

CS=Coordinated Session • IS=Invited Session
TS=Training Session • PS=Paper Session
EB=Electronic Board Session

4:05 PM-5:05 PM	Camelot (Third Floor)	EB	N1	
4:05 PM-6:05 PM	Empire Ballroom (7th Floor)	PS	N2	Equating: Small Samples and Testlets
4:05 PM-6:05 PM	Grand Ballroom (7th Floor)	PS	N3	CAT: Test Generation
4:05 PM-6:05 PM	Renaissance Ballroom (5th Floor)	PS	N4	DCM & Diagnostic Models
4:05 PM-6:05 PM	Seville Ballroom East (Lobby Level)	PS	N5	Applied International Assessment
4:05 PM-6:05 PM	Seville Ballroom West (Lobby Level)	CS	N7	Exploiting Technology in the Service of Assessment for Learning
4:05 PM-6:05 PM	Toledo (5th Floor)	CS	N8	Automated Scoring of Nontraditional Forms of Assessment
4:05 PM-6:05 PM	Valencia (Lobby Level)	PS	N9	Comparing Standard Setting Methods

CS=Coordinated Session • IS=Invited Session
TS=Training Session • PS=Paper Session
EB=Electronic Board Session