NCME

national
council on
measurement
in education

National Council on Measurement in Education

2014 Training Sessions
April 2-3

2014 Annual Meeting
April 4-6

Loews Hotel
Philadelphia, Pennsylvania

NCME Officers

President	Wim van der Linden *CTB/McGraw-Hill, Monterey, CA*
Vice President	Lauress Wise *HUMRRO, Seaside, CA*
Past President	Gregory J. Cizek The *University of North Carolina at Chapel Hill, Chapel Hill, NC*
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

Joseph Martineau
Michigan Department of Education, Lansing, MI

Cindy Walker
University of Wisconsin-Milwaukee, Milwaukee, WI

James Wollack
University of Wisconsin, Madison, WI

Huafang Zhao
Montgomery County Public Schools, Rockville, MD

Jennifer L. Kobrin, Secretary
Pearson, Wayne, NJ

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	Susan Davis-Becker *Alpine Testing Solutions, Lincoln, NE*
Website Management Committee	Brett Foley *Alpine Testing Solutions, Denton, NE*

2014 Annual Meeting Chairs

Annual Meeting Program Chairs	Paul De Boeck *The Ohio State University, Columbus, OH* Kathleen Scalise *University of Oregon, Eugene, OR*
Training and Development Committee Chair	William Skorupski *The University of Kansas, Lawrence, KS*
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 Loews Philadelphia Hotel. Stop by to pickup a ribbon and obtain your bib number for the fun run and walk. It will be open at the following times:

Wednesday, April 2	7:30 AM-4:30 PM
Thursday, April 3	7:30 AM-4:30 PM
Friday, April 4	8:00 AM-4:30 PM
Saturday, April 5	10:00 AM-4:30 PM
Sunday, April 6	8:00 AM-1:00 PM

Proposal Reviewers

Sue Bechard
Anton Beguin
Derek Briggs
Hua-Hua Chang
Chia Yi Chiu
Jerome Dagostino
Alina Davier
Paul De Boeck
John DeJong
Kristen DiCerbo
Jeff Douglas
Steve Ferrara
Mark Gierl
Ron Hambleton
Bob Henson
Matthew Johnson
Seock-ho Kim
Claudia Leacock
Won Chan Lee
John Lockwood
Ric Luecht
Krista Mattern
Andrew Maul
Marty McCall
Patrick Meyer
Andreas Oranje
Carrie Piper
Mary Pitoniak
Barbara Plake
John Poggio
Sophia Rabe-Hesketh
Frank Rijmen
Larry Rudner
Kathleen Scalise
Sandip Sinharay
Stephen Sireci
David Thissen
Gerald Tindal
Wim van der Linden
Alina von Davier
Matthias von Davier
Craig Wells
James Wollack
Ada Woo
Qing Yi
April Zenisky

Graduate Student Abstract Reviewers

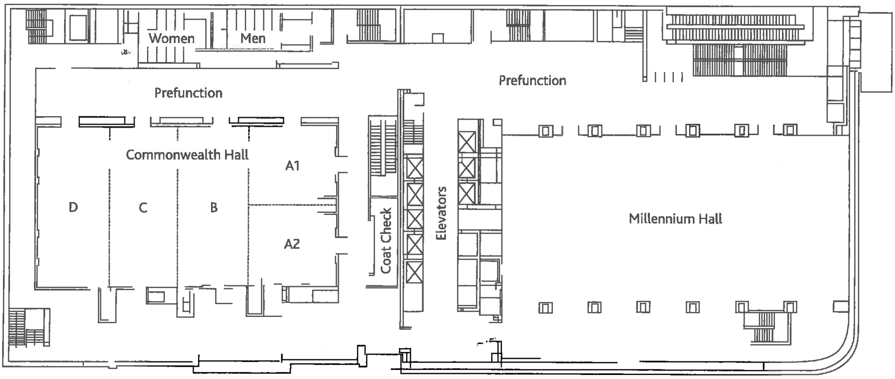
Lokman Akbay
Beyza Aksu
Katherine Allison
Allison Ames
Meagan Arrastia
Katherine Bailey
Michael Barnes
Irene Barry
Lisa Beymer
Angela Blood
Yuanchao Bo
Janine Buchholz
Kevin Cappaert
Allison Chapman
Michelle Chen
Yiling Cheng
Amy Clark
Cavwell Edwards
Anthony Fina
Fernanda Gándara
Susan Gillmor
Deborah Goins
Nese Öztürk Gübes
Yong He
Ian Hembry
Jason Herron
Xueying Hu
Sukkeun Im
Hyeonah Kang
David King
Quinn Lathrop
Isaac Li
Tuo Liu
Tanya Longabach
Xiao Luo
Xin Luo
Wenchao Ma
Tamara Miller
Kristin Morrison
Ryoungsun Park
Rachel Perlin
Surneyra Sahbaz
Sarah Scott
Hyojeong Shin
Joshua Sussman
Shuwen Tang
Ting Wang
Wei Xu
Lihong Yang
Ping Yang
Nedim Yel
Jiahui Zhang
Mingcai Zhang
Judith Zimmermann

Future Annual Meeting

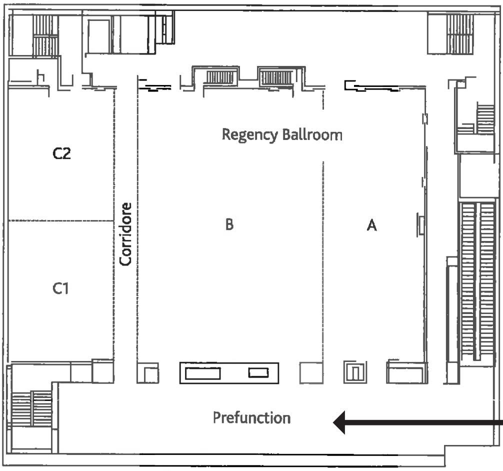
2015 Annual Meeting
April 15-19
Chicago, Illinois

Hotel Floor Plans – Loews Philadelphia

2nd Floor



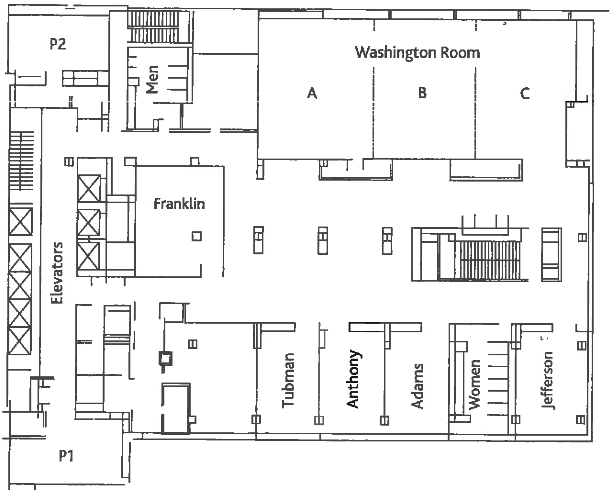
2nd Floor Mezzanine



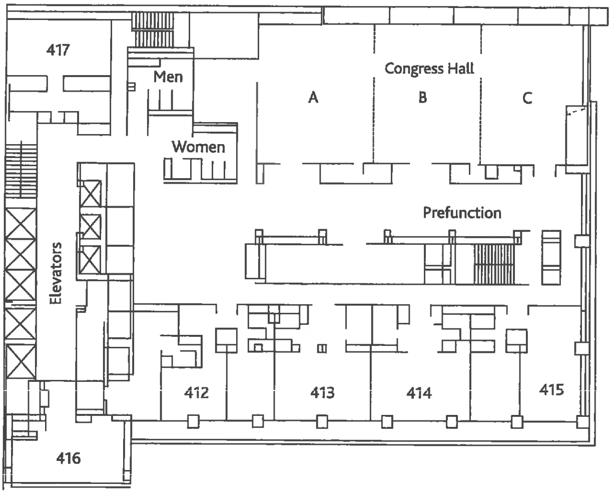
Location of NCME
Information Desk

Hotel Floor Plans – Loews Philadelphia

3rd Floor

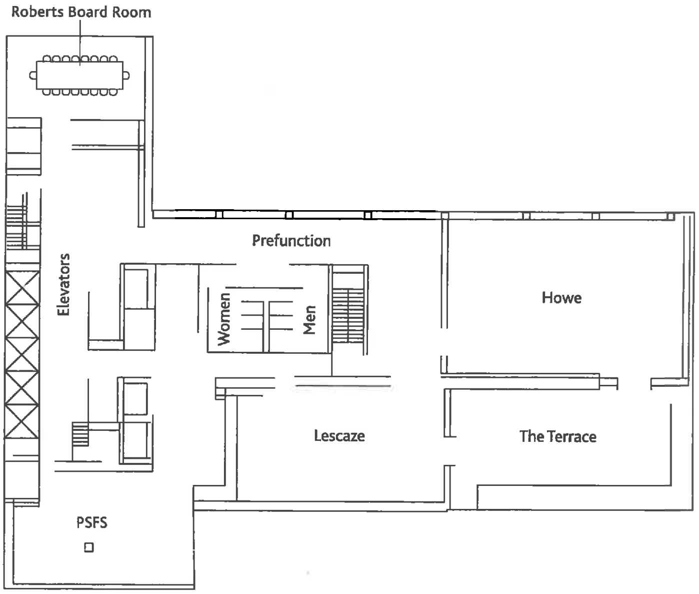


4th Floor



Hotel Floor Plans – Loews Philadelphia

33rd Floor



Pre-Conference Training Sessions

The 2014 NCME Pre-Conference Training Sessions will be held at the Loews Hotel on Wednesday, April 2, and Thursday, April 3. 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 (except for one that is scheduled from 8:30 AM-11:30 AM). 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 Loews 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.

Wednesday, April 2, 2014
8:00 AM-5:00 PM, Commonwealth B, AA

flexMIRT®: Flexible Multilevel Multidimensional Item Analysis and Test Scoring

Li Cai and Carrie R. Houts

There has been a tremendous amount of progress in item response theory (IRT) in the past two decades. flexMIRT® is a new 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 models with non-normal latent densities, and fit diagnostic classification models. This one day training session will introduce users to the flexMIRT system and provide valuable hands on experience with the software.

flexMIRT® fits a variety of unidimensional and multidimensional IRT models (also known as item factor 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. A generalized dimension reduction EM algorithm, 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. In addition to fully implementing many recently developed multilevel and multidimensional IRT models, flexMIRT supports six methods for estimating item parameter standard errors: Supplemented EM, Forward Difference, Richardson Extrapolation, empirical cross-product information matrix, Fisher (expected) information matrix, and sandwich covariance matrix. 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).

A key 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 all models implemented in flexMIRT.

....continued on page 10...

...continued from page 9...

flexMIRT has an intuitive syntax. It can import data natively in space-, comma-, or tab-delimited formats. Windows-based flexMIRT, with a friendly graphical user interface (GUI), is available in both 32-bit and 64-bit flavors. flexMIRT is designed with cross-platform compatibility and large-scale production use from the beginning. The computational engine of flexMIRT is written using standard C++, which ensures that it can run on any platform where a C++ compiler exists. 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.

flexMIRT is fast. For multidimensional and multilevel models that permit dimension reduction, flexMIRT automatically reduces the amount of quadrature computation to achieve a dramatic reduction in computational time. As modern CPU architecture trends toward multi-core design, flexMIRT uses parallel processing to further speed up computations by spreading the load automatically to the multiple available cores or processing units.

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 that could run Windows XP or above.

Wireless internet service provided.

Wednesday, April 2, 2014
8:00 AM-5:00 PM, Commonwealth D, BB

Assessing Soft Skills: K-12 to Higher Education, Domestic and International

Patrick C. Kyllonen, Richard Roberts and Jonas Bertling

Soft skills assessment in education has seen dramatic growth over the past decade. New constructs and methods are featured in international studies, including opportunity to learn, attitudes, engagement, teacher dispositions, school climate, and longitudinal noncognitive skills growth. In K-16, noncognitive assessments (e.g., ETS's Index assessment, SuccessNavigator), are used for monitoring student growth, risk, and for course placement to supplement cognitive tests. Noncognitive assessments are being used for admissions (e.g., ETS's PPI), and several noncognitive admissions experiments are underway in undergraduate, graduate, dental, business, and law schools. NCME can and should be at the forefront of this new area of assessment.

Topics include; (a) self and others' rating scales; (b) behavioral anchors and anchoring vignettes; (c) preference methods, ipsative, quasi-ipsative, and normative measurement, and IRT methods—multidimensional unfolding pairwise preference (Stark et al., 2012), and Thurstonian (Brown & Maydeu-Olivares, 2013); (d) video and textual situational judgment tests with classical and IRT scoring methods (e.g., nominal response model); (e) signal detection methods for concept familiarity; (f) methods for measuring creativity; and (g) collaborative problem solving. For each topic, a lecture reviews findings, item development, and scoring, followed by hands-on demonstrations and individual and team exercises.

We introduce participants to soft skills assessment (noncognitive, 21st century, social-emotional). We review key findings, classical and IRT-based scoring procedures, and provide practice developing, experiencing, and scoring innovative assessments.

Wednesday, April 2, 2014
8:00 AM-5:00 PM, Washington A, CC

Testing Accommodations for Computer-Administered, Postsecondary Readiness Assessments

S.E. Phillips

State and district testing accommodations policies for paper-and-pencil tests have been a work in progress for many years. Now that most states are implementing the common core content standards and assessments, focusing on college and career readiness and moving toward computer-administered assessments, state testing programs need to revise their testing accommodations policies to reflect their new goals, implementation of new technology and new uses for student test results.

Collecting relevant information, sharing perspectives with other states and applying criteria for legally and psychometrically defensible decisions before the new assessments are administered should assist state testing programs in their advance preparations. This pre-session is designed to provide structured presentations, group discussion, practice activities and teamwork strategizing about difficult testing accommodation decisions specific to participants' state testing programs.

The training session will address the following questions: what federal laws and Test Standards apply to testing accommodations decisions; what decisions have courts made about testing accommodations; how can staff obtain, summarize and apply new court decisions relevant to a particular testing program; what processes and criteria for testing accommodation decisions are most likely to produce legally/psychometrically defensible results; what are some possible alternatives for difficult testing accommodation decisions submitted by the participating testing programs?

Intended audience includes multidisciplinary teams of three to five professionals from a state, district and/or organization responsible for making testing accommodation decisions for a specific testing program. Possible team members include testing directors, content specialists, measurement specialists, policymakers (board member, legislator/legislative staff, program administrator, assistant commissioner, governor's aide), special populations staff, or representatives from a state's contractor.

Wednesday, April 2, 2014
8:00 AM-5:00 PM, Commonwealth A, DD

Diagnostic Measurement: Theory, Methods and Applications

Laine Bradshaw and Jonathan Templin

State-led assessment consortia (Partnership for Assessment of Readiness for College and Careers and Smarter Balance) have emphasized a need to design tests that efficiently provide diagnostic feedback to students, parents, and teachers. This workshop will introduce participants to a methodological framework that is useful for supporting the development of such diagnostic tests that are needed, yet lacking (Perie, et al., 2007), in large-scale testing.

Diagnostic measurement is an emerging field of psychometrics that focuses on providing actionable feedback from multidimensional tests. This workshop provides a semi-technical, hands-on introduction to the terms, techniques, and methods used for diagnosing what students know, thereby giving researchers access to information that can be used to guide decisions regarding students' instructional needs.

Overview and Objectives.
Upon completion of the workshop, participants will be able to understand the rationale and motivation for using diagnostic measurement methods. 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 remediation 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.

The target audience members are educational researchers and practitioners who are seeking to better evaluate what students know through the use of tests. This session is appropriate for graduate students, researchers, and practitioners at the emerging or experienced level. Participants are expected to have only a basic knowledge of statistics and psychometrics to enroll.

Wireless internet service provided.

Wednesday, April 2, 2014
8:00 AM-5:00 PM, Washington B, EE

Application of Principled Design and Development in Large-Scale Assessment

Kristen Huff, Sheryl Packman, Amy Hendrickson, Pamela Kaliski, Cindy Hamen, Lori Nebelsick-Gullett, Paul Nichols, Steve Ferrara, Amy Reilly, Emily Lai, and Maureen Ewing

This session will provide participants with examples of the tools, processes and outcomes of principled design and development (PDD). Presenters will discuss steps involved in PDD, and participants will engage in group work. The intended audience includes those interested in test design, item writing, validity, and aligned assessment systems.

The cornerstone of principled design and development (PDD) is an evidentiary argument requiring that each target of measurement (e.g., learning goal) for an assessment be expressed as a claim to be made about an examinee that is relevant to the specific purpose and audience(s) for the assessment. The observable evidence required to warrant each claim is also articulated. In turn, the claims and evidence shape the design of assessment opportunities for students to demonstrate what they have learned, whether that opportunity is a classroom activity or a multiple-choice item on a high-stakes assessment. Once identified, the characteristics of these assessment opportunities are referred to as task models, each capable of generating multiple assessment tasks. Taken together, the claims, evidence, and task models constitute the evidentiary argument.

The proposed training session will accomplish several goals. First, the session will provide participants with clear and detailed examples of the tools, processes and outcomes of PDD. This will comprise a thorough review of what is involved in conducting a domain analysis and domain model and in constructing the assessment framework. Participants will also be introduced to tools and guidelines for developing claims, evidence, achievement level descriptions, and task models. Throughout the discussion examples of how PDD can be applied in various contexts will be highlighted. Specific examples will be drawn from national large-scale assessments, online college assessments, and assessments for students with special needs. Importantly, the session will also cover how PDD can be applied to a formative instructional process to influence checks for understanding, design of interim assessment, and feedback. The session will illustrate how different approaches to PDD (e.g., assessment engineering) complement and differ from each other and from conventional approaches. Finally, presenters will provide an evaluation of the benefits and lessons learned when using PDD in a comprehensive assessment plan that supports instruction and learning leading up to a large-scale summative assessment. Participants will engage in hands-on work by approximating the development of a hypothetical assessment using PDD.

All presenters are qualified to provide this training given their multiple years of experience applying PDD to assessments at the College Board, Pearson, edCount and the national consortia. Specific assessments include Advanced Placement Exams, the GED exam, and a state science test.

An important theme of the training session will be how PDD provides a foundation for making stronger links between curriculum, instruction, and assessment and enhancing the test score validation argument. Understanding how principled assessment design practices can aid in the development of assessments that have the potential to be strongly aligned to curriculum and instruction is critically important and relevant given today's focus on standards-based assessment.

Wednesday, April 2, 2014
8:00 AM-12:00 noon, Commonwealth C, FF

A Graphical and Nonlinear Mixed Model Approach to IRT

Frank Rijmen and Minjeong Jeon

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 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 far lower 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 may 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 analysis). The package relies on an integration of nonlinear and generalized linear 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. The workshop will be given by Frank Rijmen and Minjeong Jeon. Frank Rijmen played a primary role in developing a nonlinear mixed model framework for item response models. More recently, he has integrated the nonlinear mixed model framework with a graphical modeling approach. He is the author of the Matlab code on which flirt is based. Minjeong Jeon has published several papers on item response theory and mixed models. She is an expert in R, and is the primary author of the R package flirt.

Wireless internet service provided.

Wednesday, April 2, 2014
8:00 AM-12:00 noon, Washington C, GG

Introduction to Natural Language Processing in Educational Measurement

Kirk A. Becker, Dmitry I. Belov, Alan D. Mead, and Bernard Veldkamp

Educational measurement practice (item bank development, form assembly, scoring of constructed response answers, test security, etc.) involves the processing of an enormous amount of text. This requires large numbers of people to write, read through, evaluate, classify, edit, score, and analyze the text. Not only is this process time consuming and resource intensive, but it is also subjective and prone to error. Subject-matter experts must define the construct of the test through some formalized process. Based on the construct, definition items are written, reviewed, edited, and classified. Beyond the individual items, item banks must also be evaluated to identify content overlap, cuing, or other content features that will lead to local item dependence and reduce construct representation. Newly written items approved for pretesting must then be administered to a sample of representative test takers before their statistical quality can be determined. If the items involve constructed response answers, they must be scored by trained human raters. Finally item writing must be conducted on a continuous basis due to security issues, and construct definition must be reevaluated on a regular basis due to changes in practice or standards.

Natural language processing (NLP) can be used to reduce the above-mentioned costs in time, money, and labor. NLP is a collection of methods for indexing, classifying, summarizing, generating, and interpreting texts. Initially, educational measurement made use of these methods in the development of automated essay scoring (AES) engines. Recently, however, NLP methods have been applied to nearly every aspect of test development and psychometrics: decision-tree-based item difficulty modeling (IDM), using text analysis to improve Bayesian priors for computerized adaptive testing (CAT) and multistage testing (MST), searching for pairs of mutually excluded items, item generating, item bank referencing, and test security.

Text classification plays a crucial role in all NLP applications for educational measurement. For example, in AES an essay may be classified into a pass or fail region; in IDM, items within a decision-tree node may be ordered by semantic similarity between their passages and keys; in test development, a group of items may be ordered from highest to lowest semantic similarity to a given item; in test security, a group of items in a brain dump can be identified by using semantic similarity measures.

This workshop will introduce participants to core concepts of NLP, with emphasis to text classification and its applications in psychometrics. Participants will study, run and modify R code demonstrating these concepts on real data sets.

Wednesday, April 2, 2014
1:00 PM-5:00 PM, Commonwealth C, HH

A Practitioner's Guide to Growth Models

Katherine Furgol Castellano and Andrew D. Ho

Growth models use longitudinal student test score data 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.

In particular, 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.

This structure and content draws primarily from the co-presenters' publication, *A Practitioner's Guide to Growth Models*, that will be available to all participants and included as part of the session fees. The co-presenters will also draw upon their previous work in growth modeling, including their recent publication in the *Journal of Educational and Behavioral Statistics*, "Contrasting OLS and Quantile Regression Approaches to Student 'Growth' Models," and their manuscript under review with the same journal that examines aggregate-level inferences of Student Growth Percentiles and other conditional status metrics. In addition, they will review and contrast alternative growth model frameworks by other researchers in this growing and fertile research area. In general, the co-presenters, Dr. Ho, an Associate Professor at Harvard Graduate School of Education, and Dr. Castellano, an IES Postdoctoral Fellow at UC-Berkeley, are dedicated to understanding growth models and their uses, limitations, and strengths as well as how to communicate this information in an understandable and actionable way to interested parties with a range of technical backgrounds.

This training session is intended for two primary audiences. The first consists of federal, state, or local education officers responsible for selecting, interpreting, estimating, and/or reporting growth model results. The second consists of researchers, including graduate students, interested in learning and developing a common framework for growth models with an emphasis on policy-relevant contrasts. Another possible audience includes those interested in conducting a course or instructional unit on growth models. Experience with simple linear regression and Excel is strongly recommended but not required. Although some session examples will use the statistical software package R, prior experience with R is not necessary.

....continued on page 18...

...continued from page 17...

Given the practical import of this session's topic and the high interest by the intended audience in understanding "what is under the hood" of growth models, we intend for this session to be highly interactive and hands-on. Our instructional methods include structured discussions, traditional presentations, and "lab-style" activities where participants actively apply growth models to real longitudinal student data. During our inaugural presentation of this training session at last year's (2013) NCME conference, we found participants were highly engaged and came prepared to ask questions pertinent to their own assessment programs or research agendas. Accordingly, for NCME 2014, we plan to contact participants prior to the session and ask them what they hope to gain from the session and what questions they would like us to answer. We will use their responses as well as our evaluations from NCME 2013 to augment our presentation materials. Given the multiplicity of topics related to growth models, this process will ensure that our session is tailored to the needs of our participants while still covering the main topics of our Practitioner's Guide to Growth Models.

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.

Wireless internet service provided.

Wednesday, April 2, 2014
1:00 PM-5:00 PM, Washington C, II

Using Visual Displays to Inform Assessment Development and Validation

Brett P. Foley

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.

Presenters will discuss the test development cycle, pointing out opportunities for the inclusion of visual displays at three major sections (i.e., design, operational development, communication of results). The session will emphasize hands-on practice with the techniques discussed. Panelists will be given step-by-step instruction in the creation of many of the visual displays discussed. Panelists will also receive relevant Excel files so they may follow along on their own laptops. 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.

Intended audience includes assessment developers, users, and consumers interested in using visual displays in assessment development and validation who have basic experience using Microsoft Excel.

Thursday, April 3, 2014
8:00 AM-5:00 PM, Commonwealth A, JJ

Analyzing NAEP Data Using Direct Estimation Approach with AM

Emmanuel Sikali and Young Yee Kim

The National Assessment of Educational Progress (NAEP) is the only nationally representative and continuing assessment of what America's students know and can do in various subject areas. Assessments are conducted periodically in mathematics, reading, science, writing, the arts, civics, economics, geography, and U.S. history. Since NAEP assessments are administered uniformly using the same sets of test booklets across the nation, NAEP results serve as a common metric for all states and selected urban districts.

NAEP assessments do not focus on individual students' achievement. The aim is to document achievement at group level. Students take different but overlapping combinations of portions of the entire pool of items. No one student receives enough test questions to provide an accurate test score. Instead, statistical methods are used to calculate score distributions for groups of students directly from answers given by many students, without the intermediate step of calculating scores of individual students. This procedure allows coverage of a large number of items in targeted content frameworks. Because of this approach, however, NAEP does not provide scores for individual participants. Instead, multiple plausible values are provided for each participant for secondary data analysts' research.

The unique features of NAEP require special considerations in analyzing NAEP data. This workshop will introduce users to the psychometric design of NAEP, sampling procedures and data analysis strategies required by these design features. These include Marginal maximum likelihood approach to computing scale scores, and appropriate variance estimation procedures. NCES released a mini-sample public-use NAEP data file in 2011. This data file provides a great opportunity to illustrate statistical analysis with NAEP data. The unique psychometric features of NAEP prohibit researchers from using common statistical software packages (e.g. SAS or SPSS) without appropriate handling. In addition, analyzing NAEP data with plausible values requires analysis to be limited to the variables included in a conditioning model to produce plausible values in examining interactions. Analyzing NAEP data using direct estimation approach allows researchers to investigate interactions among variables that were not included in the conditioning model.

Upon completion of this training seminar, it is expected that participants will:

- be familiar with the design, content, and research utility of the NAEP assessments;
- know how to use the AM software that allows users to do their own analyses;
- understand the need for using weighting and variance estimation variables correctly; and know the resources available to them at the National Center for Education Statistics (NCES) they can tap for assistance in their research.

At the start of the seminar, we will assess attendees' skills and knowledge by having them share their research interests, previous work in the area, and level of expertise in working with large data sets, complex sample design, and weighting and variance estimation procedures. This will help as we focus the lectures in each session.

We also will provide participants with electronic hard copies of all slideshow presentations and related handouts. Handouts will include descriptions of the survey instruments, examples illustrating how to use the AM software, and links to the NAEP website.

....continued on page 21...3

...continued from page 20...

The workshop will be structured to help participants meet the goals described in the summary. The draft agenda below identifies the sessions that will be held during the seminar. Throughout each of the sessions, participants will have the opportunity to reflect on how the information shared is relevant to their own research interests.

This seminar is open to advanced graduate students and faculty members from colleges and universities, and to researchers, education practitioners, and policy analysts from state and local education agencies and professional associations. Participants are expected to have working knowledge of Item Response Theory and sampling theory.

Wireless internet service provided.

Thursday, April 3, 2014
8:00 AM-5:00 PM, Commonwealth B, KK

Multidimensional Item Response Theory: Theory and Applications and Software

Lihua Yao, Mark Reckase, and Richard Schwarz

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 framework. 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. Data requirements and formats, and sample data and input files will be provided to participants prior to the workshop. Participants should bring laptop computers and any data they would like to use.

Intended audience of the session: this session is intended for upper-level graduate students, testing professionals, and researchers, who are interested in learning MIRT, MIRT linking, MCAT, and its applications, and who is working with dichotomous or polytomous data that are multidimensional in nature.

Objectives of the training session: The participants will: (a) learn the concept and gain a deep 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.

Materials to be given to participants at the session: materials include a book (*Multimensional Item Response Theory* by Mark Reckase), relevant academic papers, data and input files, and a copy of the presenters' power point presentation. Materials to be given to participants prior to the session: software and sample data can be downloaded at ww.BMIRT.com

Thursday, April 3, 2014
8:00 AM-5:00 PM, Commonwealth C, LL

Test Equating Methods and Practices

Michael Kolen and Robert L. Brennan

The need for equating arises whenever a testing program uses multiple forms of a test that are built to the same content and statistical specifications. Equating is used to adjust scores on test forms so that scores can be used interchangeably. The goals of the session are for attendees to be able to understand the principles of equating, to conduct equating, and to interpret the results of equating in reasonable ways. Equating is contrasted with related linking processes, traditional and IRT equating methodology is described, and practical issues are discussed. The focus is on developing a conceptual understanding of equating through numerical examples and discussion of practical issues.

Prior to 1980, the subject of equating was ignored by most people in the measurement community except for psychometricians who had responsibilities for equating. In the early 1980's, the importance of equating began to be recognized by a broader spectrum of people associated with testing. For example, the 1984 and 1999 AERA, APA, NCME Standards for Educational and Psychological Testing devoted a substantial part of a chapter to equating, whereas the 1974 Standards did not even list equating in the index. This increased attention to equating is attributable to at least three developments during the past 30+ years. First, there has been an increase in the number and variety of testing programs that use multiple forms of tests, and the testing professionals responsible for such programs have recognized that scores on multiple forms should be equated. Second, test developers and publishers have often referenced the role to equating in arriving at reported scores to address a number of issues raised by testing critics. Third, the accountability movement in education and issues of fairness in testing have given equating an increased emphasis among testing professionals. In addition to statistical procedures, successful equating involves many aspects of testing, including procedures to develop tests, to administer and score tests, and to interpret scores earned on tests. Of course, psychometricians who conduct equating need to become knowledgeable about all aspects of equating. The prominence of equating, along with its interdependence on so many aspects of the testing process, also suggests that test developers and all other testing professionals should be familiar with the concepts, statistical procedures, and practical issues associated with equating.

Still, our experience suggests that relatively few measurement professionals have sufficient knowledge to conduct equating. Also, many do not fully appreciate the practical consequences of various changes in testing procedures on equating, such as the consequences of many test-legislation initiatives, the use of performance assessments, and the introduction of computerized test administration. Consequently, we believe that measurement professionals need to be educated in equating methods and practices

The session is designed for upper level graduate students, new Ph.D.'s, testing professionals with operational or oversight responsibility for equating, and others with interest in learning about equating methods and practices. Participants should have at least one graduate course in measurement and two graduate courses in statistics. The co-directors authored the book T*est Equating, Scaling, and Linking: Methods and Practices*, which is used as a text for this course. The co-directors are frequently consulted by various persons throughout the country and internationally concerning equating, and they are sometimes asked to provide training sessions on equating. The co-directors have conducted an NCME training session on equating every other year since 1993.

Thursday, April 3, 2014
8:00 AM-5:00 PM, Commonwealth D, MM

Bayesian Networks in Educational Assessment

Russell Almond, Robert Mislevy, David Williamson, and Duanli Yan

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). 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: Evidence-Centered Design; Graphical Models; Graphical Modeling Tools and Applications; Refining Graphical Models with Data; and ACED: ECD in Action Demonstration.

This course is intended for people who have a good knowledge of probability and statistics (at the level of a college course in statistics with mathematics), but little experience with graphical models (Bayes nets) and related technologies.

Thursday, April 3, 2014
8:00 AM-12:00 noon, Washington B, NN

Combining Fun, Learning, and Measurement: Game Development in the Classroom and Beyond

Seth Corrigan, Shonté Stephenson, Tamas Makany, Mat Frenz, Erin Hoffman, Malcolm Bauer, Michael John, Andreas Oranje, Tanner Jackson, Chris Kitchen, Maria Bertling, Kristen DiCerbo, and Uma Menon

This half-day workshop engages participants in hands-on learning and application of a seven step agile-Evidence-Centered Design (aECD) process for designing game-based assessments. The two part workshop will introduce participants to the key steps to making functional prototypes including how to successfully navigate their production, release and continuous improvement.

This course will introduce participants to a new methodology we created that integrates Evidence-Centered Design (ECD) and the agile/scrum product development process within a game design framework (GDF) to learning games. The ECD framework details the design process of an assessment. In brief, it provides the structural framework for developing assessments from the perspective of evidentiary reasoning and validity. While participants may understand ECD for use in educational assessment or the agile process for game development, we will use the term aECD to refer to the combination of the two approaches. We will present the aECD methodology as a unified research and development approach for creating game-based formative assessment. After a brief overview of the ECD framework, GDF, and AGILE, this course will walk participants through each of the seven steps of the aECD process through discussion and hands-on activities including:

1. Greenlight - The goal of the Greenlight process is to build a coherent and well-defined business case for making the proposed game. The team also develops a working prototype of the core game mechanic based on the approach.
2. Concepting - Dedicated to creating a functional prototype that demonstrates the core mechanics and features of the game. The prototype should demonstrate a functioning and meaningful telemetry system by feeding raw data into an in-house tool.
3. Pre-Production - Dedicated to creating a vertical slice, or "first playable", fully functional slice of the game that could theoretically be deployed.
4. Production - Dedicated to expanding the vertical slice, or "first playable", into a full, feature complete game. It should be a multi-layered experience and include learning and gameplay progressions.
5. Post-Production - Mainly dedicated to developing a robust and relatively accurate back-end assessment engine.
6. Live - Dedicated to refining the underlying assessments models using the data being gathered from the unrestricted, nationwide play of the game.
7. Continuous Improvement - Dedicated to close observation of the product in the market in order to make adjustments to the assessment models, product features, and user-facing interactions based upon those observations and market feedback.

The workshop provides opportunities for participants to gain an understanding of the development and workings of game-based assessments as well as work alongside experts in game development and assessment. We expect to organize the workshop in two parts: Part I: introduce background information on game design and ECD and Part II: introduce and demonstrate how to apply 7-step aECD methodology.

This course is open to all who are interested in the methodology of Agile Evidence Centered Design (aECD) for the purpose of creating learning games. Basic understanding of ECD for uses in educational assessment is preferred, but no previous experience in game design is necessary.

Thursday, April 3, 2014
8:00 AM-12:00 noon, Washington C, OO

Effective Item Writing for Valid Measurement

Anthony Albano and Michael Rodriguez

The ability to create and critique test items is essential for educators, psychologists, researchers, test developers, and others working with educational tests. High-quality items can provide valuable information about an individual's knowledge, skills, and abilities, thereby supporting score inferences and decision-making in instructional, selection/placement, and certification/licensure settings. Item quality and the validity and usefulness of the information provided by a test item depend on a number of factors, including the appropriate use of item content (i.e., content aligned with the construct or learning targets), item type (e.g., selected-response, constructed-response), clarity of expression (e.g., correct grammar, word-usage, formatting, and structure), and cognitive load (e.g., matching the cognitive requirements of the item to the individual's capacity and ability level). Additionally, effective test items have been reviewed for sensitivity to relevant attributes and characteristics of the individual, such as ethnicity or primary spoken language; as a result, the impact of bias and construct-irrelevant variance have been reduced. These item-writing guidelines and principles are grounded in over 80 years of empirical research (Haladyna & Rodriguez, 2013), and they are the focus of this training session. The general movement in education toward higher-quality tests at all levels of instruction and with all types of learners makes this training session especially timely and of practical importance. The main goal of the session is to prepare participants to meet increasing testing-related demands in their professions by training them in item writing for effective and valid measurement.

Following the session, participants should be able to: implement empirically-based guidelines in the item writing process; describe procedures for analyzing and validating items; apply item-writing guidelines in the development of their own items; and review items from peers and provide constructive feedback based on their adherence to the guidelines.

The training session will consist of short presentations with small-group and large-group activities. Material will be contextualized within common testing applications (e.g., classroom assessment, response to intervention, progress monitoring, summative assessment, entrance examination, licensure/certification). Participants are encouraged to bring a laptop computer, as they will be given access to a web application that facilitates collaboration in the item-writing process; those participating in the session in-person and remotely will use the application to create and comment on each other's items online. This practice in item writing will allow participants to demonstrate understanding of what they have learned, and receive feedback on their items from peers and the presenters.

Wireless internet service provided.

Thursday, April 3, 2014
8:30 AM-11:30 AM, Washington A, PP

Writing Effective Research Reports

Samuel A. Livingston

The focus of the workshop will be on the writing of reports of empirical research studies and other reports that include quantitative information. Emphasis will be on the selection, sequencing, and presentation of information, not on writing style. The workshop will include three activities:

1. Panel discussion: "Good Graph, Bad Graph." Two panelists will analyze and discuss several graphs taken from actual research reports and articles. The instructor will project each graph onto a large screen for the participants to view. The panelists will point out the good and bad features of each graph and, where appropriate, suggest better ways to present the same information.
2. Illustrated lecture:"50 Ways to Write a Bad Research Report." The instructor will present a list of 50 things NOT to do when writing a research report, based on 40 years of reading and reviewing published and unpublished research reports.
3. Small-group exercise: "What's Wrong and How to Fix It." The instructor will distribute copies of a badly written research report, created especially for this exercise. This report, based on a paper written by the instructor in the 1970s, has been rewritten to incorporate many of the worst features of badly written reports. Working in groups of three to five people, the participants will read the report, identify the problems, and make a list of specific changes that would improve the report. The groups will then take turns describing one specific change they recommend: what part of the report should be changed, why it should be changed, and how it should be changed. They will continue until all the suggested changes have been presented or until time runs out.

Each participant will receive a copy of the instructor's 27 page booklet, "How to Write an Effective Research Report."

Thursday, April 3, 2014
1:00 PM-5:00 PM, Washington A, QQ

Landing Your Dream Job for Graduate Students

Deborah Harris, Nathan Wall, and Xin Li

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).

The presenters have a set of materials/slides they bring to the session, but the specific questions/ concerns of each session's attendees have shaped the specific material covered in each previously presented session. The materials provided to attendees cover more information than can be provided in the session, with the topics based on concerns observed among graduate students, particularly in the previous training sessions.

The presenters have been working in the profession for a few years to several decades. They have diverse experiences to draw on. In addition, materials from previous presenters and others interested in fostering graduate student careers are also included. Deborah J. Harris is Chief Research Scientist at ACT, Inc., and has been involved with graduate students through teaching as an adjunct at the University of Iowa, serving on dissertation committees, working with graduate Research Assistants, and working as a mentor and an organizer of the ACT Summer Intern Program. She is involved in the hiring and mentoring of staff, which frequently includes new doctorate recipients. Nathan Wall is a Research Scientist at eMetric. He has been in the educational testing business for more than ten years and completed his doctoral studies while working full time. Having worked for multiple companies during his career, he can describe to graduate students the duties that may be expected of them upon being hired. He has also worked both on site and remotely, and can speak to the differences for new professionals. Xin Li is a Research Associate at ACT, and a recent graduate from the University of Texas-Austin. She has had a variety of experiences as a graduate student, and as a relatively new professional, she can speak to finishing up her course work and dissertation, navigating the job search process, and settling into beginning a career.

Previous versions of this training session were presented in 1998, 2000-2002, 2005-2006, and in 2007 -2012. The session has been well-received by the attendees based on the feedback they have provided.

Thursday, April 3, 2014
1:00 PM-5:00 PM, Washington B, RR

Software Development Meets Statistics/Measurement Research

Damian Betebenner

In 2009, Marc Andreesson, the father of the modern web-browser through his creation of Mosaic/ Netscape in the 1990s, made that pronouncement that "software will eat the world" to emphasize the rapidly increasing importance of software (as opposed to hardware) in every aspect of life. 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. However, the expansion of the importance of software goes well beyond software packages data analysts use. Development tools alter the way that people work and collaborate. These tools are open source and build upon other open source tools allowing users to exploit computing resources and the social nature of the internet in ways impossible just a few years ago.

In this day long session, participants will be introduced to the following modern software development tools and show how, through rich, real-life working examples, they can be combined to house projects ranging from a dissertation, to a proto-type for a published articles, to a multi-state/national data analysis project.

GitHub: Version control of software is fundamental in any software developers toolkit. Version control allows for teams to work on code together in a well-coordinated fashion so that changes are well documented and integrated into the larger project. Contemporary data analysis projects often have large code bases making them ideally suited for version control. Despite this, most data analysts are not familiar with version control. GitHub, a social coding website allows for "collaboration without coordination". We currently use the service extensively for all data analysis projects associated with over 25 state growth analyses with clients to document source code publicly with strict version control.

Amazon EC2: The advent of cloud computing has "democratized" high end computing. Computation and memory intensive calculations that previously only occurred with large organizations and/ or research universities are now available to individuals for pennies. Amazon has pioneered cloud computing and we show how statistics and measurement analysts can utilize their Elastic Cloud Computing service (EC2) to carry out large scale data analyses.

Bootstrap: Bootstrap, sometimes referred to as Twitter Bootstrap is a responsive open-source web framework developed by two former employees of Twitter to facilitate the rapid development of client independent (e.g., works well on a desktop, a tablet, and a phone) high quality web content. Available in only the last couple of years, Bootstrap has become the premier framework for rapid development of high quality content. Combined with GitHub's ability to host static websites as part of projects, Bootstrap can be leveraged to produce premium quality documentation, including blogs.d3.js: d3.js (data driven documents) is a javascript based visualization framework that allows users to take data and create visualizations (interactive or non-interactive) of the highest quality. The New York Times uses this framework to produce its interactive visualizations often appearing on the front page of its website.

....continued on page 30...

...continued from page 29...

R: R is an open source statistical software environment that has become the lingua franca of statistical computing due to its flexibility in be extended and having those extensions readily shared with the larger R community. Currently there are over 5,000 user created packages available for R and its integration with tools like GitHub, EC2, and d3 is rapidly occurring.

As practicing measurement/data analysts, these tools have become an indispensable part of our work and research but are not a part of most measurement or data analyst tool kits. In this training session we will show the revolutionary impact that collaborative version control systems like GitHub can have upon the work flows practiced by measurement and data analysts.

Wireless internet service provided.

Thursday, April 3, 2014
1:00 PM-5:00 PM, Washington C, SS

TAO Workshop: A Hands-on Introduction to TAO, the Open Source Assessment Platform

Thibaud Latour

Technology is playing an ever-bigger role in assessments, offering plenty of opportunities but also some challenges. This workshop is for the technically inclined audience who would like to learn how to use computer-based assessment tools as part of their research work or their teaching. It will also help policy-makers to better evaluate the impact of computer-based testing on educational curricula – whether for K-12, Higher Ed or Professional learning.

This workshop is divided in 3 sections, covering the different parts of the assessment lifecycle: item development and item banking; test assembly and delivery; results analysis and reporting. Each section starts with a discussion of the key requirements and associated design principles, followed by a guided tour of TAO's features. By the end of this workshop, participants will be able to create simple items, prepare them for test delivery, and analyze the results.

Requirements: attendees need to bring their own notebook (Windows, Mac or Linux).

Wireless internet service provided.

Friday, April 4, 2014 • 8:00 AM - 9:40 AM, Washington
Coordinated Session, A1

Applications of Bayesian Networks in Education

Organizer and Session Chair: Russell G. Almond, Florida State University, Tallahassee, FL

Constructing Bayesian Networks for a Stealth Assessment of Qualitative Physics

Yoon Jeon Kim, Russell Almond, Valerie Shute, and Matthew Ventura, Florida State University, Tallahassee, FL

This paper describes the construction of a Bayesian network scoring model for the physics-based game Newton's Playground. The construction includes conceptualizing the network, integrating with other systems, choosing parameterizations, eliciting prior parameters, and learning parameters and refining the model using pilot data.

Using Bayes Nets in the VTG Intelligent Learning Environment

Michael J. Timms, Australian Council for Educational Research, Camberwell, Victoria, Australia

The Voyage to Galapagos web-based learning environment allows students to learn about evolution in an exploratory manner. VTG uses Bayes Nets to monitor a student's need for assistance in applying science practices and, when the probability that a student needs help reaches a threshold value, the assistance system switches on.

Determining a Reasonable Starting Place for an Instructionally Embedded Dynamic Assessment: Heuristic Versus Bayesian Network Analysis

Neal M. Kingston, University of Kansas, Lawrence, KS

To ensure reasonable initial placement in the Dynamic Learning Maps Alternate Assessment System a teacher-friendly heuristic was developed based on a school supplied survey of individual student communication and academic needs. The heuristic was validated against assessment data from 1800 students and compared with a Bayesian Network solution.

mIRT-bayes as Hybrid Measurement Model for Technology-Enhanced Assessments

Kathleen Scalise, University of Oregon, Eugene, OR and Jody Clarke-Midura, Massachusetts Institute of Technology, Cambridge, MA

A challenge in some technology-enhanced assessments (TEA) is handling complex information. A hybrid mIRT-bayes modeling approach is introduced, applied to Harvard's Virtual Performance Assessments. Results show improved precision for student proficiency estimates and improved item fit. The hybrid model appears to amplify information and improve measurement characteristics for this TEA.

Friday, April 4, 2014 • 8:00 AM - 9:40 AM, Regency A
Paper Session, A2

Constructing Adaptive Tests and Assessing Fit Characteristics

Session Chair: Yuehmei (May) Chien, Pearson, Iowa City, IA

Assessing Person Fit in a Computer Adaptive Test

John A. Stahl, Pearson, Chicago, IL

This paper explores alternative ways to detect aberrant person performance on a computer adaptive test (CAT). The alternative methods employ 1) predicted percent correct, 2) the Wald-Wolfowitz runs test and 3) regression analysis of estimated ability on sequence of item administration.

Constructing Shadow Test CAT With Variable Test Length

Qi Diao, Hao Ren, and Seung Choi, CTB/McGraw-Hill, Monterey, CA

CAT using shadow test approach currently assumes a fixed test length. This study will introduce shadow test CAT with variable-length termination. Simulation studies will be conducted to evaluate the performance of the approach in comparison to other content balancing approaches. Practical implications and future research directions will be discussed.

Impact of ICC Model Misspecification on CAT

Han Lee, University of South Carolina, Columbia, SC, Jessalyn Smith, CTB/McGraw-Hill Education, Irmo, SC, and Brian Habing, University of South Carolina, Columbia, SC

With more testing programs using computer adaptive test (CAT), it is important to understand the robustness of the algorithms implemented. The purpose of this study is to investigate, through simulation, the robustness of CAT several methods to accurately provide ability estimates when the item characteristic curves are misspecified.

An Efficient Custom-Made Adaptive Online Placement Test for Mathematics

Yuehmei Chien and Yun Jin Rho, Pearson, Iowa City, IA

An online placement testing system is introduced, which is undergoing conversion from computer-based testing (CBT) to computer-adaptive testing (CAT). This study examined the performance in terms of test length and classification precision and the gain from converting CBT to CAT in terms of shorter test length and more accurate classification.

Friday April 4, 2014 • 8:00 AM - 9:40 AM, Regency B
Paper Session, A3

Diagnostic Models: Polytomous Data and Attributes

Session Chair: Louis V. DiBello, University of Illinois at Chicago, Bloomingdale, IL

A New Diagnostic Model for Multiple-Choice Option-Based Scoring with Applications

Louis V. DiBello, Learning Sciences Research Institute, University of Illinois at Chicago, Bloomingdale, IL, Robert A. Henson, University of North Carolina at Greensboro, Greensboro, NC, and William F. Stout, Learning Sciences Research Institute, University of Illinois at Chicago, Urbana, IL

We present a new modeling framework called the Generalized Diagnostic Classification Model for Multiple Choice Option-Based Scoring (GDCM-MC). It is designed to model multiple choice assessments with misconception-linked response options while incorporating a guessing component. We describe the new model and present simulation studies and analyses of real data.

A Generalized Nonparametric Approach to Diagnostic Classification for Polytomous Attributes

Shawn Stevens, University of Michigan, Ann Arbor, MI, Robert A. Henson, The University of North Carolina at Greensboro, High Point, NC, Fu Liu, The University of North Carolina at Greensboro, Greensboro, NC, and Namsoo Shin, University of Michigan, Ann Arbor, MI

This proposal addresses an extension of the method proposed by (Chiu & Douglas, in press) to polytomous attributes. The properties of this method are studied using a simulation study and then the application of the method to study science learning progressions is emphasized.

Guessing and Multidimensional Cognitive Diagnosis: Introducing the gMLTM-D

Megan E. Lutz, Georgia Institute of Technology, Atlanta, GA

A generalization of the Multicomponent Latent Trait Model for Diagnosis (MLTM-D; Embretson & Yang, 2013) is proposed, combining the modeling of "guessing" using IRT and cognitive diagnostic modeling (CDM) approaches. Both simulated and real-world data will be used to simultaneously scale persons and items along multiple, non-compensatory components.

A Polytomous Cognitive Diagnostic Method with Weighted Distance

Shuliang Ding, Fen Luo, and Wenyi Wang, Jiangxi Normal University, Nanchang, China

A new polytomous cognitive diagnostic method named PWD is proposed. PWD combines both a cognitive model and an item response model. Simulation results show that the classification accuracy is almost over .9 for 7 attributes with PWD. Moreover a new design of test blueprint for polytomous cognitive diagnosis is introduced.

Friday April 4, 2014 • 8:00 AM - 9:40 AM, Regency C1
Paper Session, A4

Angoff Judgments, Procedures, Scores and Exercises

Session Chair: George M. Harrison, University of Hawaii at Manoa, Manoa HI

The Effect of Intrajudge Consistency Feedback in an Angoff Procedure

George M. Harrison, University of Hawaii at Manoa, Manoa, HI

Although intrajudge consistency feedback has been provided to Angoff judges in practice, little research has investigated its effect. In this randomized experiment with 36 judges, non-numerical intrajudge consistency feedback was found to significantly improve judges' accuracy with themselves. Generalizability-theory results also suggested the feedback improved interjudge consistency.

Consistency of Angoff-Based Judgments: Evidence Across Multiple Licensure Tests

Richard J. Tannenbaum and Priya Kannan, ETS, Princeton, NJ

Despite its widespread use, Angoff-based standard setting has been criticized for lack of consistency in item judgments and cutscore recommendations. We investigated consistency across nine educator licensure tests. Two independent, multistate panels were formed for each test, and multiple measures of consistency were applied. Results offer positive evidence of consistency.

The Effect of Rating Unfamiliar Items on Angoff Passing Scores

Jerome Clauser, American Board of Internal Medicine, Philadelphia, PA and Ronald K. Hambleton, University of Massachusetts at Amhert, Amherst, MA

This presentation will examine the requirement that Angoff judges provide ratings for all test items. When judges are unfamiliar with specific item content, these items may be perceived as artificially difficult. The results suggest that this requirement has the potential to suppress passing scores and ultimately undermine test validity.

The Impact of Correct Answer Indication in an Angoff Exercise

Janet Mee, Brian Clauser, Melissa Margolis, and Marcia Winward, National Board of Medical Examiners, Philadelphia, PA

This study examines the impact of explicitly indicating correct answers for items on a test with a broad content base. Results do not show that judges' ratings were impacted by providing correct answers, but the amount of time required to rate an item was reduced.

Friday April 4, 2014 • 8:00 AM - 9:40 AM, Regency C2
Paper Session, A5

Testlets, Dependence, Carry-Over

Session Chair: Chao Xie, University of Maryland, College Park, MD

Model Selection for IRT Observed Score Equating of Testlet-Based Tests

Juan Chen, Michael J. Kolen, and Deborah J. Harris, ACT, Inc., Iowa City, IA

Performance of four models on IRT observed score equating of testlet-based tests are compared. IRT observed score equating is used for 3PL IRT and GRM; unidimensional approximation of MIRT observed score equating is used for TRT and bi-factor models. Equipercentile method is used as the baseline for comparison.

Item Response Models for Carry-Over Effect in Parallel-Design Tests

Kuan-Yu Jin and Wen-Chung Wang, Hong Kong Institute of Education, Hong Kong, Hong Kong

The parallel design is commonly used in survey or inventories. Standard IRT models fail to account for carry-over effect across scales. We thus created a new of IRT model and conducted simulations to evaluate parameter recovery and consequences of model misspecification. An empirical example of school bullying is given.

Cross-Classified Modeling of Dual Local Item Dependence

Chao Xie and Hong Jiao, University of Maryland, College Park, MD

This study proposes a cross-classified model to deal with the scenario where local item dependence is caused by two cross-classified factors simultaneously. A simulation study is conducted to investigate the potential factors affecting the need to use the more complex cross-classified model over the simplified multilevel model by ignoring cross-classification.

Equating With Ducal Local Dependence Under the Anchor Test Design

Ting Xu, University of Pittsburgh, Drexel Hill, PA and Feifei Ye, University of Pittsburgh, Pittsburgh, PA

This simulation study uses concurrent calibration method under a multilevel 3PL testlet model to equate test forms in situations where test data show local item dependence and local person dependence. The results suggest that ignoring local person dependence and/or local item dependence could lead to biased person parameter estimation.

Friday, April 4, 2014 • 8:00 AM - 9:40 AM, Commonwealth A
Paper Session, A6

Standard Errors

Session Chair: Tony Thompson

Examining Scaling Methodologies and Constant CSEM for Vertical Scaling
Tony Thompson, Hongwook Suh, and J.P. Kim, ACT, Inc., Iowa City, IA

This study uses empirical data from a vertical scaling design to compare the results of a variety of scaling approaches, such as concurrent versus separate calibration, different calibration programs, and different linking methods. The study also explores the creation of a vertical scale with constant CSEM.

Asymptotic Standard Errors for True Score Equating of Polytomous Items
Cheow Cher Wong, Singapore Examinations and Assessment Board, Singapore

Building on previous works by Lord and Ogasawa for dichotomous items, this paper proposes a derivation for the asymptotic standard errors of true score equating involving polytomous items for non-equivalent groups of examinees. The proposed formulas were validated using concurrent calibration calibration equating and mean/mean equating of simulated bootstrap samples.

Conditional SEM and Reliability for Raw and Scale Scores with Multiple Item Formats
Rashid S. Almehrizi, Sultan Qaboos University, Muscat, Oman

Various researchers have proposed different distributions to estimate conditional standard error of measurement (CSEM) for scale scores for various item scoring. The paper proposes a general procedure to estimate CSEM and reliability for any scale scores for test with any item scoring resulting from using multiple item formats or stratification

An Alternative Way to Achieve Constant Conditional Standard Error of Measurement
Dongmei Li, David Woodruff, Tony D. Thompson, and Hongling Wang, ACT Inc., Iowa City, IA

A general variance-stabilizing procedure is proposed for use in equalizing the conditional standard error of measurement (CSEM) in test scaling. It is applied and evaluated in a variety of situations assuming different error distributions to illustrate its accuracy and wider applicability than the previously used arcsine transformation.

Friday April 4, 2014 • 8:00 AM - 9:40 AM, Commonwealth B
Coordinated Session, A7

Measuring "Hard-to-Measure" Aspects of Text Complexity

Organizer and Session Chair: Kathleen Sheehan, Educational Testing Service, Princeton, NJ

Measuring the Difficulty of Inferring Connections Across Sentences

Kathleen M. Sheehan and Diane Napolitano, ETS, Princeton, NJ

Many proposed cohesion metrics focus on the number and types of explicit cohesive ties detected in a text without also considering differences in the difficulty of required inferences. A new cohesion measure structured to address this limitation is proposed, and its performance as an indicator of text complexity is examined.

Automatic Detectors of Text Specificity and Organization

Ani Nenkova, University of Pennsylvania, Philadelphia, PA

We have automated the quantification of two aspects of writing. Our supervised model for sentence specificity performs robustly and accurately on newspaper texts. Our unsupervised model for text organization is driven by syntactic patterns and captures the preferred ordering of information in a domain.

Word Associations, Lexical Cohesion and Text Complexity

Michael Flor and Beata Beigman Klebanov, Educational Testing Service, Princeton, NJ

We present a new approach for measuring lexical cohesion in a text, using word association data from large scale corpora. The novel measure, Lexical Tightness, strongly correlates with grade level of reading materials (r=-0.5 to -0.6): simpler texts are tight (more lexically cohesive) and complex texts are less cohesive.

Automatic Idiom Recognition

Anna Feldman and Jing Peng, Montclair State University, Montclair, NJ

One of the evaluative criteria for text complexity is the use of figurative language, especially, metonymy, metaphors and idioms. Our work addresses automatic detection of idioms in text. The main goal of this research project is to develop a language independent method for automatic idiom recognition.

Friday April 4, 2014 • 8:00 AM - 9:40 AM, Commonwealth C
Paper Session, A8

Aberrant Item Responses

Session Chair: J. Michael Clark, Pearson, Tulsa, OK

Comparing IRT Proficiency Estimators for Aberrant Responding and Multistage Testing

Sooyeon Kim and Tim Moses, Educational Testing Service, Princeton, NJ

The purpose is to determine which IRT proficiency estimators are more robust than others in terms of recovering examinees' true proficiency and routing accuracy when aberrant responses are present. Various aberrant responses are imposed on the simulated data of a 2-stage MST to investigate the effectiveness of IRT proficiency estimators.

Collusion Detection by Divergence with Jointly Modeled Response Times

Anne Thissen-Roe and Michael S. Finger, Comira, San Mateo, CA

Belov's (2013) divergence method detects test collusion by individuals, nested within communicating groups, via person-fit of an item response model. Here, his protocol is extended to use the van der Linden (2007) hierarchical framework for joint modeling of responses and response times. Conditions supporting added value are investigated.

The Impact of Aberrant Response Patterns on Response Similarity Statistics

J. Michael Clark, Pearson, Tulsa, OK

Researchers have developed numerous methods to identify data manifestations of unique forms of misconduct, but little attention has been given to interactions among manifestations and methods. This paper presents the results of a study investigating the impact of response pattern aberrance on the performance of a common response similarity statistic.

Methods to Detect Group-Level Aberrance in State Standardized Assessment

Feifei Li, Lixiong Gu, and Venassa Lall, Educational Testing Service, Princeton, NJ

This study examines some methods to detect group-level aberrance based on analyses of test scores and item responses respectively. The methods on test scores include regression model and large score change across years. The methods on item responses include regression model, DIF analysis, and K-index.

Friday April 4, 2014 • 8:00 AM - 9:40 AM, Commonwealth D Paper Session, A9

Multidimensionality Issues

Session Chair: Carolyn J. Anderson, University of Illinois, Champaign, IL

Log-Multiplicative Association Models: Multidimensional Item Response Models for Polytomous Items

Carolyn J. Anderson, University of Illinois, Champaign, IL and Hsiu-Ting Yu, McGill University, Montreal, QC, Canada

Log-multiplicative association models can be used as multidimensional IRT models for polytomous items. LMA models are fit to response patterns and parameters estimated using Newton-Raphson without specifying the marginal distribution of theta(s). We provide theoretical derivations of LMA models as MIRT models and demonstrate the empirical performance of the models.

Consequences of Estimating Unidimensional Item Parameters Across Multidimensional Test Forms

Ki L. Matlock, University of Arkansas at Fayetteville, Fayetteville, AR and Ronna Turner, University of Arkansas, Fayetteville, AR

Not all test developers maintain equal length and difficulty within sub-content areas across multiple forms. The purpose of this study is to investigate the effects of estimating unidimensional item parameters across multidimensional datasets with confounding length and difficulty with dimensions. Over- and underestimation of parameters were severe on unbalanced forms.

Using Graphical Multidimensional IRT Models in Large-Scale Educational Assessments

Jesper Tijmstra, Utrecht University, Utrecht, Netherlands and Frank Rijmen, Educational Testing Services, Princeton, NJ

As large-scale educational assessments increasingly focus on assessing fine-grained skills, it is becoming crucially important to correctly and efficiently model the effects of item clustering. This paper proposes the use of graphical models, which provides a flexible and computationally efficient approach to dealing with multidimensionality.

Consequences of Multidimensional Scaling on Equating Conversions

Michael Chajewski, The College Board, New York, NY, Anita Rawls, The College Board, Newtown, PA, and Jennifer Lord-Bessen, Fordham University, New York, NY

Using the Multidimensional Item Response Theory Three-Parameter Logistic Model (M3-PL) true score equating is conducted under three varying dimensional scaling alternatives. Consequences are explored in terms of dimensional stability, scaling maintenance and equating consistency across three years.

Friday, April 4, 2014, 9:40 a.m.-10:00 a.m.
Regency Ballroom Foyer

Refreshment Break

Friday April 4, 2014 • 10:00 AM-11:40 AM, Washington
Invited Session, B1

21st Century Skills Debate

Moderator: Patrick Kyllonen, Educational Testing Service

Panel Members and Presenters:

Critical Thinking is Both Less and More Than You Might Think

Nathan Kuncel, Department of Psychology, University of Minnesota

Critical thinking is not a single universally important skill. When people point to critical thinking they actually mean one of two things: a set of independent skills used in evaluating arguments that are not of universal importance at work or very domain specific expertise. Each of these two types of "critical thinking" have very different implications for policy and interventions.

The Influence of High-Stakes Testing on Instructional Practices: Are Testing Programs Supporting or Impeding the Acquisition of 21st Century Skills?

Mariale Hardiman, School of Education, John Hopkins University

Within the field of education, "21st Century Skills" has become synonymous with students' ability to demonstrate critical thinking and creative problem-solving. Can assessments measure those skills while evaluating the complex work of teachers? In an era of high-stakes accountability, testing policies drive teaching practices. Can testing improve teaching or simply devour instructional time with little benefit to the learning process?

Friday April 4, 2014 • 10:00 AM - 11:40 AM, Regency A
Invited Session, B2

A Look at Our Psychometric History: Contributions of International Scholars

Organizers and Session Chairs: Sandip Sinharay, CTB/McGraw-Hill, Monterey, CA and Linda Cook, Educational Testing Service, Princeton, NJ

Contributions of Georg Rasch

Susan Embretson, Georgia Institute of Technology, Atlanta, GA

Georg Rasch was a Danish mathematician, statistician, and psychometrician famous for the development of a class of measurement models including a treatment of dichotomous data known as the Rasch model. He also proposed generalizations and extensions of such models, and worked on other issues in statistics.

Contributions of Roderick McDonald

Hariharan Swaminathan, University of Connecticut, Storrs, CT

Professor Roderick McDonald has contributed greatly to the fields of factor analysis, structural equations models, nonlinear factor analysis, and item response theory. This presentation highlights Professor McDonald's influence on the field, focusing on his accomplishments as a person and as a psychometrician.

Contributions of Karl Jöreskog

Irini Moustaki, London School of Economics

Joreskog brought together two traditions: factor analysis and simultaneous equation modelling. The combination of the two became known as linear structural equation modelling (SEM) and opened exciting possibilities for the analysis of social data. SEM as we know it today has been shaped by Joreskog's theoretical, methodological and computational contributions.

Contributions of Louis Guttman

Charles Lewis, Educational Testing Services, Princeton, NJ

Louis Guttman (1916 – 1987) was an Israeli citizen who was born in New York and died in Minneapolis. Here, three of his many important foundational contributions to quantitative models in the social sciences will be discussed. Specifically, these contributions are in the areas of reliability, factor scores and Guttman scaling.

Friday April 4, 2014 • 10:00 AM - 11:40 AM, Regency B
Coordinated Session, B3

Scoring Related Challenges and Solutions in Technology-Enabled Assessments

Organizer and Session Chair: Mo Zhang, Educational Testing Service, Princeton, NJ

Generalizability in Automated Essay Scoring - An In-Depth Analysis

Mo Zhang and Paul Deane, Educational Testing Service, Princeton, NJ

This presentation concerns the generalizability issues of automated scoring models for essay items. We will share our research findings and provide a thorough discussion on the meaning of generalizability in automated scoring context. The suitability and scalability of implementing automated scoring capability in technology-based writing assessments will also be discussed.

Automated Scoring of an Extended Mathematics Task Involving Quadratic Functions

James H. Fife, Educational Testing Service, Princeton, NJ

This presentation will explore how an extended task involving quadratic functions was created to allow most of the items in the task to be scored automatically. In addition, the results of a pilot study will be presented, including an item analysis and a discussion of frequent incorrect responses.

Automated Scoring in Game-Based Assessments

Diego Zapata-Rivera, Educational Testing Service, Princeton, NJ

This presentation will discuss the application of various automated scoring capabilities, especially the ones presented in this symposium, in game-based assessments. Eight validity issues when designing game-based assessments will be elaborated, and the challenges and opportunities of using automated scoring as well as the validation concerns will also be discussed.

Short Answer Scoring in an Innovative Test Design

Michael Heilman and Paul D. Deane, Educational Testing Service, Princeton, NJ

This presentation explores how short-answer scoring technologies play out in the context of an experimental English Language Arts assessment system, developed as part of a research initiative exploring the use of online scenario-based assessments with a strong cognitive foundation. In particular, we will present and evaluate short-answer scoring models based upon machine learning algorithms.

Friday April 4, 2014 • 10:00 AM - 11:40 AM, Regency C1
Paper Session, B4

Topics in Measurement Gaps: The Need to Know

Session Chairs: Benjamin Shear, Stanford University, Stanford, CA and Katherine E. Castellano, Educational Testing Service, San Mateo, CA

An Analysis of Norms for Early Reading Curriculum-Based Measures

Gerald Tindal, Jessica Saven, Joseph F.T. Nese, Daniel Anderson, University of Oregon, Eugene, OR; and Joe Betts, Riverside Publishing, Rolling Meadows, IL

In this presentation we describe curriculum-based measurement norms in early reading skills for students in Grades Kindergarten and 1. We present normative distributions at three time periods within the year on various measures and use hierarchical generalized linear modeling with student characteristics and slope to predict risk in the fall.

The Kids We Cannot Match When Measuring Student Growth

Adam E. Wyse, Michigan Department of Education, Arden Hills, MN and John Denbleyker, Minnesota Department of Education, St. Paul, MN

The use of growth models in state accountability systems is widespread. This study shows that certain groups of students were less likely to be matched when applying student growth models. These student groups tended to represent groups that are of greatest concern to policy makers when measuring student growth.

Simultaneous Estimation of Multiple Achievement Gaps From Ordinal Proficiency Data

Katherine E. Castellano, University of California at Berkeley, Berkeley, CA; Benjamin R. Shear, Sean F. Reardon, Stanford University, Stanford, CA; and Andrew D. Ho, Harvard University, Cambridge, MA

Test score summaries often only report frequencies of students scoring in a small number of ordinal proficiency categories. Using these limited data, we present a method to simultaneously estimate achievement gaps between two or more groups.

The Riddle of Dimensional Structure in Surveys of Teacher Preparation

Derek Briggs, University of Colorado, Boulder, CO

Empirical data is used to to show that a bifactor IRT model can serve as a compromise between EFA and CFA approaches to establishing the dimensional structure of a survey instrument. The presentation also focuses on the importance of replicating evidence in support of dimensional structure with new data.

Friday April 4, 2014 • 10:00 AM - 11:40 AM, Regency C2
Paper Session, B5

Model Fit Statistics / Propensity Score Matching

Session Chair: Litong Zhang, CTB, Monterey, CA

Limited-Information Methods to Assess Overall Fit of Diagnostic Classification Models

Daniel P. Jurich, James Madison University, Harrisonburg, VA, Laine P. Bradshaw, University of Georgia, Athens, GA, and Christine E. DeMars, James Madison University, Harrisonburg, VA

Lack of practical methods to assess model fit presents a crucial issue in diagnostic classification modeling (DCMs). This study analytically illustrates how a recently proposed limited-information model fit framework can be extended to evaluate DCMs. Simulation and real data analyses were conducted to investigate performance of these indices.

A Scaled F-Distribution to Approximate the SEM Test Statistic Distribution

Hao Wu, Boston College, Chestnut Hill, MA and Johnny Lin, Educational Testing Service, Princeton, NJ

When the assumption of multivariate normality is violated in structural equation models, the null distribution of the test statistic is not χ^2 distributed. Methods exist to approximate the statistic under violations, but can only match up to two moments. We propose using a scaled F-distribution to match three moments simultaneously.

Using Propensity Score Matching to Examine Interrupted Students' Scores

Litong Zhang, CTB, Monterey, CA, Dong-In Kim, CTB, Denver, CO, Ping Wan, CTB, Chesterfield, Bin Wei, CTB, Chicago, IL, Ying Feng, CTB, Princeton, NJ, Jessalyn Smith, CTB, Columbia; Sandra McGuire, and Sara Kendall, CTB, Monterey, CA

During spring 2013 summative testing windows, interruptions to online tests occurred in several states. This study uses propensity score matching methods to create equivalent student groups in order to determine what impact the test interruptions had on student scores. Several matching methods and impacted students groups will be compared.

Friday April 4, 2014 • 10:00 AM-11:40 AM, Commonwealth A
Invited Session, B6

Joint NATD and NCME Symposium

Session Chair and Organizer:
Bonnie Swan, NATD Vice President/President-Elect, University of Central Florida

Moderator:
Steve Addicot, Vice President, Caveon Test Security

Panel Members:
Rachel Schoenig, Head of Test Security ACT
Walt Drane, Office of Student Assessment, Mississippi Department of Education
James Wollack, Director of Testing Services, University of Wisconsin—Madison
John Fremer, President, Caveon Consulting Servcies

Trustworthy Test Results—Cheating in the Past, Present, and Future

This symposium will provide a comprehensive understanding of past and current issues, and the crucial test security issues our industry must face in the not-too-distant future. Each panel member will explore a test-security related topic—either something you have contended with already, or one you are likely to encounter in the near future.

Friday April 4, 2014 • 10:00 AM - 11:40 AM, Commonwealth B Paper Session, B7

Measuring Digital Learning Across Countries and Over Time

Organizer and Session Chair: John Ainley, ACER, Camberwell, Australia

Measuring Changes in ICT Literacy Over Time

John Ainley, Julian Fraillon, Wolfram Schulz, and Eveline Gebhardt, ACER, Camberwell, Australia

Measuring change over time is especially challenging in the field of ICT Literacy where changes in technologies change the contexts in which proficiencies are assessed. This paper reports an approach to measuring change in ICT Literacy over three assessment cycles spanning six years at Grades 6 and 10.

Assessing Computer and Information Literacy Across 20 Countries

Julian Fraillon, Australian Council for Educational Research, Melbourne, Australia

This presentation describes the process of developing, the substantive content and the measurement properties and content of an achievement scale based on the responses of 8933 students from 20 countries to the International Computer and Information Literacy (ICILS) field trial test items.

Reading in PISA: From Print to Digital

Juliette Mendelovits, Australian Council for Educational Research, Camberwell, Australia

This paper explores country and gender differences in performance on digital and print reading in PISA 2009 and 2012 and offers some explanations for these differences related to text structures and features of texts, and the concomitant skills and understandings that readers exercise in navigating print and digital texts.

Assessing Learning in Digital Networks, Perspectives from ATC21S

Mark R. Wilson, University of California at Berkeley, Berkeley, CA, Kathleen Scalise, University of Oregon, Eugene, OR, and Perman Gochyyev, University of California, Berkeley, Berkeley, CA

Learning in digital networks is an important 21st century skill. This paper discusses a construct that has been conceptualised around four strands and associated descriptions of progress. It describes the authentic tasks developed to capture assessment data and the results of analyses of pilot studies conducted in four countries.

Friday April 4, 2014 • 10:00 AM - 11:40 AM, Commonwealth C
Paper Session, B8

Reliability Issues

Session Chair: Steve Culpepper, University of Illinois at Urbana-Champaign, Flossmoor, IL

A Repeated Attempts Model for Computerized Cognitive Assessments

Steve Culpepper, University of Illinois at Urbana-Champaign, Flossmoor, IL

An IRT model is developed for cases when test-takers are allowed repeated attempts on items. Students with higher latent scores tend to have more knowledge and are better able to retrieve that knowledge in fewer attempts. Estimation strategies and simulation results are included. The methodology is applied to classroom assessments.

The Nonparametric Approach to the Glb Bias Correction

Wei Tang and Ying Cui, University of Alberta, Edmonton, Canada

The Glb is theoretically the optimal lower bound estimate of reliability. Yet its sampling bias hinders its application, thus statistical treatment of the bias is badly needed. This study proposed and evaluated the nonparametric methods for correcting the Glb sampling bias.

A Principled Approach to Designing Reliability Studies

Paul Nichols, Pearson, Iowa City, IA, Emily Lai, Pearson, Iowa City, IA, and Jefferey Steedle, Pearson, Austin, TX

Identification of facets in a reliability study appears to be based on subjective judgment. We present evidence suggesting an audience may accept estimates of reliability but reject that conditions are replications. We describe using principled assessment to identify facets and conditions so reliability estimates reflect explicit and defensible criteria.

Cronbach's Alpha as an Improper Linear Model: Is There a "Robust Beauty" Here Too?

Emily Ho, Fordham University, Bronx, NY; Walter Kaczetow, and Jay Verkuilen, CUNY Graduate Center, New York, NY

Cronbach's alpha is problematic as a reliability estimate due to its downward bias. Through simulations, we hypothesize that re-conceptualizing alpha in an improper linear model framework with unit weights will yield more robust and consistent estimates across samples, compared to other reliability coefficients (i.e. McDonald's omega) which contain sampling error.

Friday April 4, 2014 • 10:00 AM - 11:40 AM, Commonwealth D
Paper Session, B9

Multidimensional Models for Polytomous Data

Session Chair: Lisabet M. Hortensius, University of Minnesota, Minneapolis, MN

Practical Guidelines for the Estimation of Multidimensional Ordered Polytomous Models

Lisabet M. Hortensius and Chun Wang, University of Minnesota, Minneapolis, MN

Estimating multidimensional ordered polytomous model parameters involves choosing between models (generalized partial credit model and graded response model), estimation methods, and software. The goal of this study is providing practical guidelines for making these choices based on a simulation study exploring estimation accuracy using three software packages on several datasets.

Pseudo-Likelihood Estimation of Multidimensional Polytomous Item Response Theory Models

Youngshil Paek and Carolyn J. Anderson, University of Illinois at Urbana-Champaign, Champaign, IL

Multidimensional item response theory models (MIRT) are presented as log-multiplicative association (LMA) models, which do not require numerical integration to fit to data. However, standard MLE of LMA models is limited to small numbers of items. We propose and study a flexible pseudo-likelihood algorithm that overcomes limitations of MLE.

A Noncompensatory Multidimensional IRT Model for Multiple-Choice Items

Sien Deng and Daniel Bolt, University of Wisconsin at Madison, Madison, WI

A multidimensional noncompensatory model for multiple-choice items is presented. The model assumes response selection occurs through stepwise process, whereby different skills apply to different steps. The model is applied to a sentence correction test (N=5000), for which such a response process appears plausible, and multidimensionality is confirmed.

A MIRT Approach to Measuring Response Styles in Large-Scale-Assessment Data

Lale Khorramdel and Matthias von Davier, Educational Testing Service, Research and Development, Princeton, NJ

A multidimensional item response theory (MIRT) approach to measure and correct for response styles (RS) is presented. It is shown that RS can be measured unidimensional and differentiated from trait-related responses. Results are discussed with regard to possible cultural differences and correlations between noncognitive scales and cognitive test scores.

Friday April 4, 2014 • 11:50 AM-12:40 PM, Washington
Invited Session, C1

Session Chair: Kathleen Scalise, University of Oregon, Eugene, OR

Measuring Learner Engagement with Data Mining

Ryan Baker
Teachers College, Columbia University

In recent years, there has been increasing interest in measuring learner engagement in automated ways. In this talk, I discuss my group's work to develop and validate measures of a range of engaged and disengaged behaviors, as well as affect, within student use of online learning, using a combination of field observation and data mining.

Friday April 4, 2014 • 11:50 AM-12:40 PM, Regency B
Invited Session, C2

Session Chair: Paul De Boeck, Ohio State University, Columbus, OH

True Grit

Angela Duckworth
University of Pennsylvania

Who succeeds in life? Is it simply those among us who are gifted and talented in our pursuits? Or, is aptitude merely "potential" if effort and interest are not sustained over time? In this presentation, Angela Duckworth reviews her research on grit, the tendency to pursue challenging goals over years with perseverance and passion. She describes the predictive power of grit for performance in the National Spelling Bee, graduation from West Point, graduation from the Chicago Public Schools, and a variety of other contexts. Next, she describes current work on the underlying motivational, cognitive, and behavioral mechanisms explaining what makes gritty individuals different from others. Finally, she summarizes her current thinking on how we can cultivate grit in ourselves and in others.

Overall Electronic Board Session Chairs: Marc Julian, Data Recognition Corp., Athens, GA and Mary Pitoniak, ETS, Princeton, NJ

Friday, April, 4, 2014 • 12:50 PM - 2:20 PM, Millennium Hall
Electronic Board Paper Session 1

Bayesian Networks

Session Chair: Jinnie Choi, Rutgers, The State University of New Jersey, New Brunswick, NJ

Electronic Board Presentation #1.1A (Monitor 1)
Validation of a Mathematics Learning Map Using Bayesian Network Analysis
Angela Broaddus, University of Kansas, Lawrence, KS, Zach Pardos, University of California at Berkeley, Berkeley, CA; Zachary Conrad and Ayse Esen, University of Kansas, Lawrence, KS

This study will investigate the validity of a learning map used to guide the development of a cognitive diagnostic assessment measuring understanding of the connection between fractions and division. Results from Bayesian network analyses will be used to evaluate and refine the theory depicted in the learning map.

Electronic Board Presentation #1.2A (Monitor 2)
A Comparison of General Diagnostic Models and Bayesian Networks
Haiyan Wu, National Board of Osteopathic Medical Examiners, Chicago, IL and Russell Almond, Florida State University, Tallahassee, FL

This study examines the similarities and differences between GDMs and Bayesian networks using both simulated data and real data sets. Their performances in data generation and estimation in various conditions, following the ECD framework, are investigated. Several indices for accuracy and precision in item and person levels are reported.

Electronic Board Presentation #1.1B (Monitor 1)
Bayesian Networks for Skill Diagnosis and Model Validation
Hua Wei, Pearson, Plymouth, MN, Yun Jin Rho, Pearson, Boston, MA, and John Behrens, Pearson, Austin, TX

This study evaluates how well Bayesian networks diagnose specific subskills in the domain of arithmetic with fractions, and how different domain models compare with each other in terms of the extent to which they accurately predict students' mastery of different subskills.

Electronic Board Presentation #1.2B (Monitor 2)
Bayesian Network Analysis of Contingencies in Learning Progressions
Jinnie Choi, Rutgers, The State University of New Jersey, New Brunswick, NJ, Ji Seung Yang, The University of Maryland, College Park, MD, and Ravit G. Duncan, Rutgers, The State University of New Jersey, New Brunswick, NJ

We use Bayesian networks (BN) to investigate how students develop understanding of multiple related constructs simultaneously over the course of an instructional unit in genetics. We use real data to illustrate the construction and evaluation of BNs. Results will suggest ways to model complex interplay of constructs in learning progressions.

Friday, April, 4, 2014 • 12:50 PM - 2:20 PM, Millennium Hall
Electronic Board Paper Session 2

Classification Models and Approaches

Session Chair: Margarita Olivera-Aguilar, Educational Testing Service, Princeton, NJ

Electronic Board Presentation #2.1A (Monitor 3)
Parameterizing Covariates Affecting Attribute Classification in Multiple-Groups RDINA
Yoon Soo Park, University of Illinois at Chicago, Chicago, IL, Taeje Seong, Korea Institute for Curriculum and Evaluation, Seoul, Republic of Korea, and Young-Sun Lee, Teachers College, Columbia University, New York, NY

Covariate extension of the Multiple-Groups Reparameterized DINA (MG-RDINA) model is proposed, which provides group-specific attribute parameters. Empirical analyses of six countries from TIMSS 2007 4th grade mathematics using LatentGOLD show country-specific attribute distributions for gender and science ability. Simulation results indicate stability in parameter recovery and utility of the model.

Electronic Board Presentation #2.2A (Monitor 4)
A Mixture IRT Model Analysis in Exploring Kindergarteners' Mathematical Strategy
Young-Sun Lee, Teachers College, Columbia University, New York, NY, Yoon Soo Park, University of Illinois at Chicago, Chicago, IL, and Seung Yeon Lee, Ewha Womans University, Seoul, Republic of Korea

Mixture IRT models have become a common technique to filter heterogeneous populations into homogeneous subpopulations, where examinee behavior can be more specifically analyzed. The goal of this paper is to examine whether there were differences in strategy use in mathematics problems between American and Korean kindergarteners using mixture IRT models.

Electronic Board Presentation #2.1B (Monitor 3)
Approaches to Classify Examinees in a Multidimensional Space
Liyang Mao and Xin Luo, Michigan State University, East Lansing, MI

This study described and compared four classification procedures: the non-compensatory disjunctive cut, the full-compensatory cut, the partially-compensatory cut, and the non-compensatory conjunctive cut. The impact of correlation among dimensions on classification was also investigated. The implications of different classification procedures on educational assessments were discussed.

Electronic Board Presentation #2.2B (Monitor 4)
Using Latent Profile Analysis to Identify Profiles of College Dropouts
Margarita Olivera Aguilar, Ross Markle, and Steven Robbins, Educational Testing Service, Princeton, NJ

Latent Profile Analysis will be used to examine the profiles of students in community colleges. The profiles will be determined by psychosocial, demographic and academic preparation variables. Further, using a multi-group approach the profiles of students returning to the second semester of college will be compared to non-returning students.

Friday, April, 4, 2014 • 12:50 PM - 2:20 PM, Millennium Hall
Electronic Board Paper Session 3

DIF Identification, Causes, and Consequences

Session Chair: Holmes Finch, Washington State University, Pullman, WA

Electronic Board Presentation #3.1A (Monitor 5)
Examining the Accuracy of Confirmatory Mixture Models in DIF Detection
Kelli M. Samonte, University of North Carolina at Greensboro, Walkertown, NC; John Willse, and Randy Penfield, University of North Carolina at Greensboro, Greensboro, NC

The current study aims to examine the accuracy of confirmatory mixture models in DIF detection. Past research cautions against the use of exploratory mixture IRT models in DIF contexts. Constraints placed on the traditional mixture IRT models may provide some benefits to class membership and parameter recovery in DIF contexts.

Electronic Board Presentation #3.2A (Monitor 6)
When DIF Purification Fails: A Study of Small Scales
Ronna C. Turner, Wen-Juo Lo, Elizabeth Keiffer, and Ki Matlock, University of Arkansas, Fayetteville, AR

Multi-step purification procedures ameliorate the differential impacts of unbalanced DIF contamination on inflated type I error and power for the minority group and deflated type I error and power for the majority group when analyzing large achievement-type scales. This study provides feedback on limitations when attempting purification with small scales.

Electronic Board Presentation #3.1B (Monitor 5)
DIF Analysis in Multilevel Data
Yao Wen and Cindy M. Walker, University of Wisconsin at Milwaukee, Milwaukee, WI

A simulation study will be conducted to identify the level-2 DIF in multilevel data using multilevel Rasch models and multilevel mixture Rasch models. The location of DIF, the magnitude of DIF, the number of DIF items at level 2 will be considered in the study.

Electronic Board Presentation #3.2B (Monitor 6)
Recursive Partitioning to Identify Potential Causes of Differential Item Functioning
Holmes Finch, Maria Finch, Ball State University, Muncie, IN; and Brian French, Washington State University, Pullman, WA

Model-based recursive partitioning was employed to investigate and identify potential causes of differential item functioning (DIF). Results illustrate that this approach can be an effective tool for illuminating the nature and causes of DIF by finding distinct patterns of DIF in the population.

Friday, April, 4, 2014 • 12:50 PM - 2:20 PM, Millennium Hall Electronic Board Paper Session 4

DIF: Specifics on Power, Purification, Guessing, and Criteria

Session Chair: Christine E. DeMars, James Madison University, Harrisonburg, VA

Electronic Board Presentation #4.1A (Monitor 7)
Purification of the Matching Criterion in Equated Pooled Booklet Method
HyeSun Lee, University of Nebraska-Lincoln, Lincoln, NE and Kurt F. Geisinger, University of Nebraska-Lincoln, Buros Center for Testing, Lincoln, NE

This research investigated effects of purification of the matching criterion on the equated pooled booklet method with the Mantel-Haenszel DIF procedure in large-scale assessments. Four manipulated factors were examined in a simulation study. The issues on the purification and the length of anchor were addressed.

Electronic Board Presentation #4.2A (Monitor 8)
A Power Formula for SIBTEST
Zhushan Li, Boston College, Chestnut Hill, MA

A power formula for the SIBTEST procedure for differential item functioning (DIF) is derived. It provides a means for calculating the sample size in planning DIF studies with SIBTEST. Factors influencing the power are discussed. The correctness of the power formula is confirmed by simulation studies.

Electronic Board Presentation #4.1B (Monitor 7)
Detecting DIF Using Empirical Criteria for Logistic Regression
Vincent Kieftenbeld and Seung W. Choi, CTB/McGraw-Hill Education, Monterey, CA

Likelihood ratio tests based on logistic regression generally detect DIF with good power but inflated type I error rates. One potential solution to this problem is to use empirical criteria derived from parametric bootstrapping or permutation tests. We investigate the performance of these methods and illustrate with empirical data.

Electronic Board Presentation #4.2B (Monitor 8)
Modeling DIF With the Rasch Model: Group Impact and Guessing
Christine E. DeMars and Daniel P. Jurich, James Madison University, Harrisonburg, VA

Using the Rasch model to estimate DIF when there are large group differences in ability and correct guessing leads to misestimation of DIF. Difficult non-DIF items appear to favor the focal group and easy non-DIF items appear to favor the reference group. Effect sizes for DIF items are biased correspondingly.

Friday, April, 4, 2014 • 12:50 PM - 2:20 PM, Millennium Hall
Electronic Board Paper Session 5

Aspects of Student Growth Estimation

Session Chair: Ronli Diakow, New York University, New York, NY

Electronic Board Presentation #5.1A (Monitor 9)
A Model for the Relationship Between Prior Achievement and Growth
Ronli Diakow, New York University, New York, NY

This paper discusses a model relevant to research about whether interventions alleviate or exacerbate existing differences in achievement. The model incorporates different conceptions of prior knowledge and their interaction with growth for students receiving different instruction. The presentation emphasizes application and interpretation using empirical data from an efficacy study.

Electronic Board Presentation #5.2A (Monitor 10)
Reporting Student Growth Percentiles With Estimated Standard Errors of Measurement
Jinah Choi, Won-Chan Lee, Catherine Welch, and Stephen B. Dunbar, University of Iowa, Iowa City, IA

This paper presents a procedure for estimating conditional standard errors of measurement, constructing confidence intervals, and computing reliability for Student Growth Percentiles based on binomial error assumptions. The reporting and interpretation of the measurement error of SGPs derived from longitudinal data, and the practical implications for test users are discussed.

Electronic Board Presentation #5.1B (Monitor 9)
The Effect of Conditioning Years on the Reliability of SGPs
Craig Wells, Stephen Sireci, and Louise Bahry, University of Massachusetts at Amherst, Amherst, MA

The valid application of student growth percentiles (SGPs) is partly dependent on the estimates having a reasonably small magnitude of random error. The purpose of the present study was to examine the effect of the number of conditioning years on the reliability of SGPs.

Electronic Board Presentation #5.2B (Monitor 10)
Understanding Growth and College Readiness With a Growth Mixture Model
Anthony D. Fina, Stephen B. Dunbar, and Timothy N. Ansley, The University of Iowa, Iowa City, IA

The relationship between student growth and college outcomes is investigated. Over ten years of student data (state test scores, high school and college records) are used to describe the relationship between student performance, growth, and college outcomes. Growth mixture models are used to evaluate latent classes and predict future outcomes.

Friday, April, 4, 2014 • 12:50 PM - 2:20 PM, Millennium Hall
Electronic Board Paper Session 6

Score Reports: What is New on the Horizon?

Session Chair: Mary Roduta Roberts, University of Alberta, Edmonton, Alberta, Canada

Electronic Board Presentation #6.1A (Monitor 11)
A Language Parents Understand: Reporting and Adult Functional Literacy Research
Margaret Anderson, University of Kansas, Lawrence, KS

Communicating performance information to stakeholders is a long standing issue for the measurement field. Individual student performance reports prepared under NCLB were evaluated based on a rubric from adult functional literacy (White, 2010) to identify recommended practices when developing reports for diverse audiences (e.g., parents and teachers).

Electronic Board Presentation #6.2A (Monitor 12)
Examining Score Reports Through a Communications Frame of Reference
Mary Roduta Roberts, University of Alberta, Edmonton, Alberta, Canada, Chad Gotch, Washington State University, Pullman, WA, and Jessica Nina Lester, Indiana University at Bloomington, Bloomington, IN

Informed by a conversation analysis framework, this study investigates the function of language in student score reports. Results may provide for improved score report functioning within the larger enterprise of teaching, learning, and assessment; better foundations for stakeholder engagement; and consideration of additional roles for measurement professionals.

Electronic Board Presentation #6.1B (Monitor 11)
Interactive Score Reports for State Assessments: Practices and Directions
MinJeong Shin, Xi Wang, University of Massachusetts at Amherst, Amherst, MA; Katrina Roohr, Educational Testing Service, Princeton, NJ; Abdolvahab Khademi, Molly Faulkner-Bond, Fernanda Gándara, April L. Zenisky, Stephen G. Sireci, University of Massachusetts at Amherst, Amherst, MA; and Dan Murphy, Pearson, Hadely

This study investigates current practices used by state education agencies for online score reporting by visiting SEA websites and reviewing online score reports for statewide educational assessment programs, and provides recommendations for developing interactive online databases for consumers of the results of educational tests such as parents and teachers.

Friday, April, 4, 2014 • 12:50 PM - 2:20 PM, Millennium Hall
Electronic Board Coordinated Session 7

Subscore Reporting in Adaptive Testing

Organizer: Wim van der Linden, CTB/McGraw Hill, Monterey, CA
Session Chair: Daniel Lewis, CTB/McGraw Hill, Monterey, CA

Electronic Board Presentation #7.1A (Monitor 13)
Robustness of MCAT Subscore Recovery With Varying Item Pool Structures
Haskell Sie, Pennsylvania State University, State College, PA; Keith Boughton, Seung Choi, CTB/McGraw-Hill, Monterey, CA; Lihua Yao, Defense Manpower Data Center, Seaside, CA, and Litong Zhang, CTB/McGraw-Hill, Monterey, CA

Four MCAT item selection methods were compared for the same item pool calibrated differently: all items simple-structured, all items complex-structured, and some items are simple- and others complex-structured. Recovery of the domain abilities was compared across different methods and pool structures.

Electronic Board Presentation #7.2A (Monitor 14)
Subscale Reporting in CAT Using the Augmented Subscore Approach
Xin Luo, Michigan State University, East Lansing, MI, Qi Diao, and Hao Ren, CTB/McGraw-Hill, Monterey, CA

Two methods of augmenting subscores in the context of CAT and three approaches to CAT subscore estimation are presented: unidimensional Bayesian, augmented unidimensional Bayesian, and multidimensional estimates. The results indicate that augmentation can improve subscore accuracy. However, for short tests, subscoring does not add much extra value to the total scores.

Electronic Board Presentation #7.1B (Monitor 13)
Subscore in Cat Using Higher-Order and Hierarchical IRT Models
Moonsoo Lee, UCLA, Los Angeles, CA; Xinhui Xiong, and Seoung Choi, CTB/McGraw-Hill, Monterey, CA

To report reliable and valid overall scores and sub-domain scores in CAT, the performances of four multidimensional IRT methods were compared: (a) simple-structure multidimensional IRT, (b) higher-order IRT, (c) bifactor IRT, (d) and a two-tier model. The feasibility and effectiveness of those four models were investigated under various conditions.

Electronic Board Presentation #7.2B (Monitor 14)
Subscoring By Sequencing an Adaptive Testing Battery
Fu Liu, University of North Carolina at Greensboro, Greensboro, NC; Jie Li, and Seung Choi, CTB/McGraw Hill, Monterey, CA

This study evaluates the impact of the adaptive sequencing of the subtest in an adaptive test battery on the estimated subscore and proposes how to calculate the total scores based on the high-order ability structure. A simulation study was conducted to investigate the impact of test length and covariance ability structures on all score estimates.

Friday, April, 4, 2014 • 12:50 PM - 2:20 PM, Millennium Hall
Electronic Board Coordinated Session 8

Validity Arguments for Measures of Teacher Effectiveness

Organizer and Session Chair: Jennifer Merriman, The College Board, Newtown, PA

Electronic Board Presentation #8.1A (Monitor 15)
A Framework to Support the Validation of Educator Evaluation Systems
Erika Hall, Center for Assessment, Iowa City, IA

Educator evaluation systems are based upon complex assumptions about the appropriateness of system measures and the mechanism by which desired goals will be achieved. This paper provides a framework to support validation by outlining key design assumptions within the context of a clearly defined theory of action.

Electronic Board Presentation #8.2A (Monitor 16)
A Validity Framework for a Scientific Knowledge for Teaching Assessment
Jennifer Merriman, The College Board, Newtown, PA, Steve Sireci, University of Massachusetts at Amherst, Amherst, MA, Andrew Wiley, Alpine, Oren, UT, and Cindy Farrar Hamen, The College Board, New York, NY

In this paper we present a validity argument for a Scientific Knowledge for Teaching assessment for AP Biology teachers. We tie the framework to a theory of action and lay out the propositions, claims, and evidence associated with the assessment to develop a research agenda.

Electronic Board Presentation #8.1B (Monitor 15)
Evaluating the Validity of Student Learning Objectives
Katie H. Buckley, Harvard University, Cambridge, MA

States and districts have incorporated SLOs into their educator evaluation systems; however, evidence is needed to substantiate the claims to support the use of SLOs in these systems. We present a validity argument for SLOs and evidence from one state's evaluation system, to determine if SLOs are operating as intended.

Electronic Board Presentation #8.2B (Monitor 16)
Validating a Measure of Teacher Formative Assessment Readiness
Anne Davidson, CTB/McGraw Hill, Carson City, Christina Schneider, CTB, Columbia, and Jessalyn Smith, CTB/McGraw Hill, Irmo, SC

This study will explore the relationships among elementary mathematics teachers' skill in (a) determining what an item measures, (b) analyzing student work, (c) providing targeted feedback, and (d) determining next instructional steps and investigate an instrument designed to measure teacher readiness to engage in formative assessment.

Friday, April, 4, 2014 • 12:50 PM - 2:20 PM, Millennium Hall
Electronic Board Paper Session 9

Value-Added Models, Methods and Analysis

Session Chair: Michelle C. Croft, ACT, Inc., Iowa City, IA

Electronic Board Presentation #9.1A (Monitor 17)

Handling Missing Data in Value-Added Models – An Empirical Study

Ming Li, University of Maryland, Vienna, VA; Tongyun Li and Robert W. Lissitz, University of Maryland, College Park, MD

This empirical study aims at investigating how different ways of handling missing data affect teacher effect estimates. Listwise deletion, multiple imputation and maximum likelihood methods are applied before the covariate adjustment model is used for the analysis. Teacher effect estimates following the three methods are then compared.

Electronic Board Presentation #9.2A (Monitor 18)

Impact of Approaches to Handling Dropout on Value-Added Analysis

Paula Cunningham, University of Iowa, Grinnell, IA; Catherine Welch, and Stephen Dunbar, University of Iowa, Iowa City, IA

Several approaches to handing missing data were evaluated for their effects on value-added estimates. Pairwise deletion, regression imputation, and multiple imputation produced rankings that were very highly correlated. The study suggests pairwise deletion is best where imputed data cannot be used, but regression or multiple imputation are superior otherwise.

Electronic Board Presentation #9.1B (Monitor 17)

Investigating Reliability of Value-Added Scores for Teacher Evaluation

Chih-Kai Lin, Jinming Zhang, University of Illinois at Urbana-Champaign, Champaign, IL; Furong Gao, and Daniel Lewis, CTB/McGraw-Hill, Monterey, CA

The current study aims to empirically investigate the reliability of value-added scores for teacher evaluation across 11 different value-added models (VAMs). Results suggest that the rank ordering of these value-added scores is consistent across the different models. Implications are discussed in light of caveats in applying VAMs in teacher evaluation.

Electronic Board Presentation #9.2B (Monitor 18)

Will Courts Shape Value-Added Methods for Teacher Evaluation?

Michelle C. Croft and Richard Buddin, ACT, Inc., Iowa City, IA

This presentation examines legal challenges to teacher evaluation procedures. Some lawsuits would limit the use of value-added assessments (Cook v. Stewart), while others would require the use of student achievement in evaluations (Vergara v. California). The shifting legal environment will require states and districts to reassess teacher evaluation methods.

Friday, April, 4, 2014 • 12:50 PM - 2:20 PM, Millennium Hall Electronic Board Coordinated Session 10

Experimental Research on Test Item Design and Construction

Organizer and Session Chair: James Masters, Pearson-VUE, Powell, OH

Electronic Board Presentation #10.1A (Monitor 19)
Manual and Automatic Manipulation of Test Item Distractors
Mark Gierl, University of Alberta, Edmonton, Alberta, Canada

This presentation combines research projects from several programs experimentally evaluating the generation of test items, specifically the evaluation and functioning of test item distractors. By studying the content of good versus bad distractors, and documenting the effect that changes to those distractors has on item performance, we improve our ability to create good items.

Electronic Board Presentation #10.2A (Monitor 20)
Comparing Content-Parallel Items of Different Item Formats
Joy Matthews-Lopez, National Association of Boards of Pharmacy, Athens, OH and Paul E. Jones, National Association of Boards of Pharmacy, Spanish Fork, UT

This study compares performance of multiple-choice items to strictly parallel, alternative item types on a high-stakes CAT licensure exam. "Twin" items of different formats were marked as enemies and simultaneously piloted on a population of licensure examinees. Alternative items were more informative, more difficult, and more discriminating than MC counterparts.

Electronic Board Presentation #10.1B (Monitor 19)
Measurement Characteristics of Content-Matched Medical Diagnosis Items in Different Formats
David Swanson and Kathleen Z. Holtzman, NBME, Philadelphia, PA

The series of studies described in this presentation experimentally compared the measurement characteristics (item difficulty/discrimination, response times, efficiency) of content-matched test material varying in presentation format (independent items vs matching sets) and number/source of options. Results indicated statistically and practically significant effects on difficulty, response time, and efficiency.

Electronic Board Presentation #10.2B (Monitor 20)
Effects of Item Type on Item Statistics and Timing Data
Matthew J. Burke, AICPA, Yardley, PA; Joshua Stopek and Henrietta Eve, AICPA, Ewing, NJ

This presentation discusses research studies performed to gather data comparing different item types in testing the skills of a professional field. Several operational and psychometric questions will be addressed with data from studies designed to compare item types varying in authenticity relative to the tasks of the profession.

Friday April 4, 2014 • 2:30 PM - 4:10 PM, Washington Coordinated Session, D1

Meeting the Challenges to Measurement in an Era of Accountability

Organizer and Session Chair: Henry Braun, Boston College, Chestnut Hill, MA

Psychometric Considerations for Alternative Assessments and Student Learning Objectives

Suzanne Lane, University of Pittsburgh, Pittsburgh, PA and Charles DePascale, National Center for the Improvement of Educational Assessment, Dover, NH

This paper addresses psychometric issues in the design of alternative assessments and SLOs used in accountability systems for non-tested subjects. Comparability, reliability, generalizability, and classification accuracy, each are viewed in terms of the conflict between the instructional and measurement goals with a focus on the validity of the intended inferences.

Designing Authentic Assessments to Serve Accountability Uses

Scott F. Marion, Center for Assessment, Dover, NH and Katie Buckley, Harvard University, Arlington, MA

This paper explores the challenges of designing authentic assessments for high stakes accountability uses. We discuss how authentic assessment designs may differ depending on the specific accountability use and articulate both general and specific design principles, depending on the accountability model, paying special attention to the influence of Campbell's Law.

End of Course Tests for Student Assessment and Teacher Evaluation

Denny Way, Pearson, Iowa City, IA and Steve Ferrara, Pearson, Washington, DC

We evaluate the suitability of end of course tests, designed to produce information about student achievement, for use in teacher evaluation models. We organize this evaluation around the Principled Questions (PQ) Process, which focuses on the interpretations and uses that test information users intend to make about examinees.

Psychometric Considerations for End of Course Exams

Leah McGuire, Michael Nering, and Pasquale DeVito, Measured Progress, Dover, NH

End of Course (EOC) exams have become a popular alternatives to comprehensive assessments. EOCs require specific psychometric considerations due to their narrow scope and variety of purposes. Psychometric concerns related to EOCs are discussed in this study, using examples from operational EOC exam programs for illustration.

Friday April 4, 2014 • 2:30 PM - 4:10 PM, Regency A
Paper Session, D2

Scoring and Exposure in Adaptive Settings

Session Chair: Prathiba Natesan, University of North Texas, Denton, TX

Scoring Penalty for Incomplete Computerized Adaptive Tests

Lianghua Shu, CTB/McGraw-Hill, Seaside, CA

If the not-reached items were treated as wrong, examinees might randomly guess remaining items right before out of test time. This is to find a penalty procedure to avoid any coachable test-taking strategy and to encourage examinees to answer items they have time to consider in the Computerized Adaptive Tests (CAT).

Item Selection and Exposure Control in Unidimensional Calibration of Multidimensionality

Prathiba Natesan, Kevin Kalinowski, and Robin Henson, University of North Texas, Denton, TX

We compared three item selection methods, alone, as well as in combination with Sympson-Hetter (SH) exposure for 2-dimensional data calibrated as unidimensional 3-PLMs. The results suggest that existing unidimensional CAT implementations might consider employing stratification methods, rather than recalibrate the item pool under a multidimensional model.

New Statistical Test for CAT Item Drift and Compromised Items

Fanmin Guo, Graduate Management Admission Council, Reston, MD

This paper presents a new statistic method for detecting drift in item difficulty for CAT programs. It compares and statistically tests the difference between the expected and observed proportions of correct answers from the same group of examinees responding to an item. It also applies to linear tests.

Using Speed-Accuracy Scoring Rules in Adaptive Testing

Usama Ali, ETS, Princeton, NJ and Peter van Rijn, ETS Global, Amesterdam, Netherlands

The availability of response time (RT) information in computer based testing brings the opportunity for studying various forms of using such information. We present scoring rules and derive appropriate psychometric models. These scoring rules account for both accuracy and speed, and we investigate their use for computerized adaptive testing.

Friday April 4, 2014 • 2:30 PM - 4:10 PM, Regency B
Coordinated Session, D3

Three Ways to Improve Subscores Using Multidimensional Item Response Theory

Organizer and Session Chair: Lan Huang, University of Minnesota, Minneapolis, MN

Improving the Reliability of Differences Between Pairs of Subscores

Mark L. Davison, Lan Huang, University of Minnesota, Minneapolis, MN; Christie Plackner and Sherri Nooyen, Data Recognition Corporation, Maple Grove, MN

One way to look at an examinee's strengths and weaknesses is to compare subscores across content domains. An orthogonal MIRT model will be proposed to improve the reliability of differences between subscore pairs. Comparisons with unidimensional and simple structure MIRT models will also be discussed.

Examining Subscore Reliability Within the Multidimensional IRT Framework

Okan Bulut, American Institutes for Research, Washington, DC

The lack of reliability of subscores is an important issue that might take away from the diagnostic value of subscores. This study introduces a new reliability approach as a measure of between-person and within-person variation, and compares unidimensional and multidimensional IRT subscore estimates based on this approach.

A Restricted Bi-Factor Model of Subdomain Relative Strengths and Weaknesses

Yu-Feng Chang, Qin Zhang, Shiyang Su, and Luke Stanke, University of Minnesota, Minneapolis, MN

This study proposes a bi-factor model with a general dimension score measuring overall ability and specific dimension scores which constitute a within person description of strengths and weaknesses. Each score has an accompanying conditional standard error. Results from achievement data will be used to illustrate model fits, score reliability, and interpretation.

Three MIRT Methods to Improve Subscores: Similarities and Contrasts

Mark Reckase, Michigan State University, East Lansing, MI

This presentation will describe similarities (e.g. extensive use of item parameter constraints), differences (e.g. interpretations of subscores), and limitations (e.g. possible estimation problems) of the methods for improving subscores described in the session. Implicit in all three presentations is an assumption that well-chosen item parameter constraints can improve subscores.

Friday April 4, 2014 • 2:30 PM - 4:10 PM, Regency C1
Paper Session, D4

Solutions in the Development of Technology-Enhanced Items

Session Chair: Roberta N. Chinn, PSI Services, LLC, Folsom, CA

Challenges Facing Content Experts When Developing Technology-Enhanced Items

Tim Hazen, Iowa Testing Programs, Iowa City, IA

Challenges observed in three areas-- item authoring, item review, and item analysis – during the development by content experts of technology-enhanced (TE) items are discussed. Data and case studies from ongoing large scale TE development efforts help describe these challenges and inform the solutions offered.

Automatic Item Generation Implemented for Measuring Artistic Judgment Aptitude

Nikolaus Bezruczko, Measurement and Evaluation Consulting, Chicago, IL

An automatic item generation algorithm was developed to improve objectivity and validity of artistic judgment aptitude testing. Cognitive information processing model established theoretical foundations for manipulating complexity and redundancy in an image model. Templates were produced and isomorphs are currently operational. Components and measurement analysis provide substantial validity support.

Designing Cognitive Structures to Facilitate Development of Key Features Problems

Roberta N. Chinn, Folsom, CA and Norman R. Hertz, Progeny Systems Corporation, Manassas, VA

Cognitive modeling structures were developed to facilitate development of case based scenarios (Key Features Problems) requiring candidates to identify the subset of options which are most essential to patient care. Implications for design of item models and use with technology based item generation methods are discussed.

Friday April 4, 2014 • 2:30 PM - 4:10 PM, Regency C2
Paper Session, D5

Compromised Items and Invalid Data

Session Chair: Joseph Rios, University of Massachusetts at Amherst, Amherst, MA

Examining Erasure Behavior to Investigate Compromised Items Under Multilevel Settings

Xin Li, Chi-Yu Huang, Tianli Li, and Hongling Wang, ACT, Inc., Iowa City, IA

This study examined how different erasure analysis methods perform in detecting potential test irregularities when data were generated with students nested within classrooms and imposed with erasure behavior. Various factors (i.e., classroom size, degree of classroom dependency, erasure pattern) were considered and included in the study.

Identifying Unmotivated Examinees in Low-Stakes Assessment: A Comparison of Approaches

Joseph A. Rios, University of Massachusetts at Amherst, Northampton, MA; Ou Lydia Liu, and Brent Bridgeman, Educational Testing Service, Princeton, NJ

Self-report (SRE) and response time effort (RTE) procedures were compared for identifying unmotivated examinees in a low-stakes assessment. RTE was found to provide: 1) higher correlations with test performance, 2) more efficient identification of unmotivated examinees, and 3) more effective filtering of invalid data compared to SRE.

Detection of Test Compromise With Gained Scores and Generalized Residuals

Yuming Liu, ETS, Pennngton, NJ and Yuli Zhang, ETS, Princeton, NJ

Applied Segall's (2002) IRT model for test compromise and Haberman's (2011) pairwise generalized residual method to physics and math tests of a large scale international testing program. Model parameters were estimated with Winbugs and residuals were estimated with Haberman's MIRT software. Validity evidence of the methods was examined.

Utilizing Response Time in Sequential Detection of Compromised Items

Edison Choe, University of Illinois at Urbana-Champaign, Champaign, IL

To improve detection of compromised items in CAT, a sequential procedure is proposed in which significant changes in the rate of correct responses as well as the average response time are statistically monitored over repeated administrations. Simulations are conducted to evaluate the Type I error and power of the procedure.

Friday April 4, 2014 • 2:30 PM - 4:10 PM, Commonwealth A
Paper Session, D6

Difficulties in Realizing Equivalence and Equating

Session Chair: Anthony Albano, University of Nebraska, Lincoln, NE

Maintaining Equivalent Classification Decisions Across Forms with Small Samples

Andrew C. Dwyer, Castle Worldwide, Inc., Morrisville, NC and Diane M. Talley, University of North Carolina at Chapel Hill, Chapel Hill, NC

This study examines the effectiveness of four approaches for maintaining equivalent classification decisions across test forms with small samples: circle arc equating, identify equating, resetting the performance standard, and rescaling the performance standard. The latter has received almost no attention in the literature but is showing promise.

Establishing Linking for Newly Implemented Assessments with Highly Heterogeneous Samples

Shuhong Li, Jiahe Qian, and Brown Terran, Educational Testing Service, Princeton, NJ

The sample of a newly implemented assessment with multiple test forms often consists of groups with heterogeneous abilities, demographics, and volatile sample sizes. This paper intends to develop an optimal and practically viable mechanism for establishing stable linkings for assessments in their initial stage.

Equating Mixed Format Tests Using Limited Common Item Set

Seohong Pak, University of Iowa, Coralville, IA and Won-Chan Lee, University of Iowa, Iowa City, IA

This study evaluated the feasibility of equating mixed format test using MC common items only. Four factors (sample size, effect size for new form, correlation between MC and FR, and common item length) were considered in simulating the conditions for traditional equating methods.

Implementing and Evaluating Equating With Small-Scale Assessments

Anthony Albano, University of Nebraska, Lincoln, NE

Equating methods have traditionally been developed and applied in the context of high-stakes, large-scale testing. This study provides guidelines for equating in small-scale situations, such as with assessments used to make formative decisions at the classroom level. Applications with real data are demonstrated and discussed.

Friday April, 4, 2014 • 2:30 PM - 4:10 PM, Commonwealth B Coordinated Session, D7

Integrating Games, Learning, and Assessment

Organizer and Session Chair: Andreas Oranje, ETS, Princeton, NJ

Do Game-Based Assessments Keep Their Promises? Evaluating Principles and Effects

Geneva Haertel, Terry P. Vendlinski, and Harold Javitz, SRI International, Menlo Park, CA

Criteria and procedures are identified that were used to determine the success of game-based assessment. The criteria and procedures included documentation of the Evidence-Centered Design (ECD) process used in game design and outcomes of a series of validation studies.

How Do Game Design Frameworks Align With Learning and Assessment Design Frameworks?

Erin Hoffman, GlassLab, Richmond, VA; Michael John, and Tamas Makany, GlassLab, Redwood City, CA

This paper addresses the parallels in game, learning, and assessment design through the lens of GlassLab's findings from developing SimCity: Pollution Challenge. Topics include the distinction between developing original educational games and adapting existing software; design-oriented understanding of a game's mechanical affordances; approaches to scaffolding game mechanics to expose skills.

Looking for Evidence: Telemetry and Data Mining in Game-Based Assessments

Kristen DiCerbo, Pearson, Avondale, AZ; Shonte Stephenson, and Ben Dapkiewicz, Institute of Play, Redwood City, CA

The use of fine-grained log file data to make inferences about players' knowledge, skills, and attributes requires the extraction of information that can be used in the creation of measurement models. Techniques of exploratory data analysis and educational data mining can be used to uncover patterns that identify such evidence.

Flexibility, Adaptivity, and Measurement: Psychometric Paradigms for Game Design

Shonte Stephenson, GlassLab, Redwood City, CA; Robert J. Mislevy, Andreas Oranje, Maria Bertling, and Yue Jia, ETS, Princeton, NJ

Designing valid game-based assessment raises many challenges: getting the content right and targeting the right levels of skills. We present three psychometric paradigms that provide a starting point for connecting assessment to game design frameworks and offer comments on what statistical machinery connect to each of the three paradigms.

Friday April 4, 2014 • 2:30 PM - 4:10 PM, Commonwealth C
Paper Session, D8

Anchoring and Equating

Session Chair: Anton A. Béguin, Cito Institute for Educational Measurement, Arnhem, Netherlands

Application and Justification of Vertical Comparison Instead of Vertical Linking

Anton A. 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.

Robustness of Anchor Item Flagging Criteria in IRT Equating

Ying Lu and Hongwen Guo, Educational Testing Service, Princeton, NJ

Anchor items play a crucial role in determining the quality of the equating when a common-item nonequivalent groups equating design is used. The purpose of this study is to examine the robustness of flagging criteria used for anchor item evaluation in IRT common item equating.

The Effect on Equating of Anchor Test Structure and Length

Joseph Fitzpatrick and William P. Skorupski, University of Kansas, Lawrence, KS

The equating performance of three different anchor test designs—minitests, miditests, and semi-miditests—is evaluated using simulated data and four different anchor test lengths. Results are discussed in terms of comparability with the typical minitest design, as well as the trade-off between equating accuracy and flexibility in test construction.

Evaluation of Linking Items in Vertical Scale Construction

Joshua Goodman, Pacific Metrics, Jenks, OK and Minh Duong, Pacific Metrics, Monterey, CA

This study examines the effects of using different methods to screen and remove linking items in vertical scale construction using a common item design and concurrent multi-group IRT calibration when the assumption of unidimensionality holds to varying degrees.

Friday April 4, 2014 • 2:30 PM - 4:10 PM, Commonwealth D
Paper Session, D9

Longitudinal Models

Session Chair: James S. Roberts, Georgia Institute of Technology, Atlanta, GA

Measuring and Explaining Both Individual-Level and Cluster-Level Change

Sun-Joo Cho, Vanderbilt University, Nashville, TN

It is common to collect longitudinal item response data from individuals within clusters. This paper demonstrates use of an explanatory multilevel longitudinal item response model to explain changes at both the individual-level and cluster-level. An example of the model is provided to evaluate an instructional intervention.

Marginal Bayesian Estimation in the Multidimensional Generalized Graded Unfolding Model

Vanessa M. Thompson and James S. Roberts, Georgia Institute of Technology, Atlanta, GA

This paper develops a marginal maximum a posteriori method to estimate item parameters in the multidimensional generalized graded unfolding model. Person parameters are estimated with an expected a posteriori technique. A simulation study demonstrates good parameter recovery, and the model is applied to data from a large abortion attitude survey.

An Explanatory Longitudinal Multilevel IRT Approach to Instructional Sensitivity

Alexander Naumann, Jan Hochweber, and Johannes Hartig, German Institute for International Educational Research (DIPF), Frankfurt am Main, Germany

We propose an explanatory longitudinal multilevel item response model to evaluate items' instructional sensitivity by (a) quantifying sensitivity depending on students' classroom membership and (b) relating sensitivity to instructional measures (teaching quality and content coverage). Results suggest that the model works well in its application to empirical data.

Measurement Error Correction for the Student Growth Percentile Model

Yi Shang, John Carroll University, University Heights, OH, Adam Van Iwaarden, University of Colorado, Boulder, CO, and Damian Betebenner, National Center for the Improvement of Educational Assessment, Dover, NH

This study examines measurement error correction in the Student Growth Percentile (SGP) model. Results from Monte Carlo experiments suggest that individual SGP should not be corrected, and that aggregated class SGP benefit from measurement error correction when there is student sorting. Analysis of actual assessment data confirms the experiment results.

Friday April 4, 2014 • 4:05 PM-6:05 PM
AERA Annual Meeting
Philadelphia Convention Center, 200 Level, Room 204A

Standards for Educational and Psychological Testing: Major Changes and Implications to Users

Chairs: Wayne Camara and Suzanne Lane

Rationale for the Revision of the Standards and Charge to the Joint Committee
Wayne Camara and Suzanne Lane

Major Changes in Response to the Charge
Laurie Wise and Barbara Plake

Validity and Fairness
Linda Cook and Frank Worrell

Reliability and Scaling
Mike Kane and Mike Kolen

Workplace Testing
Laurie Wise

Educational Testing
Denny Way and Barbara Plake

Program Evaluation and Policy
Brian Gong

Session Participants:
Lauress L. Wise, Human Resources Research Organization
Barbara S. Plake, University of Nebraska at Lincoln
Frank C. Worrell, University of California at Berkeley
Linda L. Cook, Educational Testing Service
Brian Gong, National Center for the Improvement of Educational Assessment, Inc.
Michael T. Kane, ETS
Michael J. Kolen, University of Iowa
Denny Way, Pearson
Suzanne Lane, University of Pittsburgh
Wayne T. Camara, ACT

Friday, April, 4, 2014 • 4:20 PM - 6:00 PM, Washington
Paper Session, E1

Featured Contributions

Session Chair: Matthew S. Johnson, Teachers College, Columbia University, New York, NY

How to Fake It: Considerations for Implementing Simulations in Psychometrics

Richard A. Feinberg, National Board of Medical Examiners, Philadelphia, PA and Jonathan D. Rubright, American Institute of Certified Public Accountants, Ewing

Simulations are widely used in psychometrics, yet the motivation for choosing this methodology is seldom articulated. This paper explains the theoretical rationale for simulations and summarizes how they are utilized to answer questions that other methods cannot. Practical recommendations for simulation designs and the presentation of simulation results are given.

A Statistical Model Framework for Test Equating

Jorge Gonzalez, Department of Statistics, Pontificia Universidad Catolica de Chile, Santiago, Chile

Statistical models are used to learn about the data generating mechanism and are described using probability models in which data are viewed as realizations of random variables. We elaborate on the idea of viewing equating as standard statistical models by introducing formal terminology as is used in defining statistical models

Evaluating the NAEP Statistical Model With Posterior Predictive Checks

Matthew S. Johnson, Teachers College, Columbia University, New York, NY and Sandip Sinharay, CTB, Monterey, CA

This presentation will evaluate the assumptions of the NAEP model with Bayesian posterior predictive checks. We examine discrepancy measures to evaluate the assumptions about test dimensionality, the simple structure of items, and assumptions about omitted and not reached items, which are treated as incorrect and missing at random respectively.

Multinomial Propensity Score Matching: An Application on PISA Outcomes

Mustafa Yilmaz, University of Kansas Center for Educational Testing and Evaluation, Lawrence, KS

This study examines the question "How successful would students from top ten countries in PISA be if they were in the USA?" by conducting a multinomial propensity score matching. I have obtained significantly different results compared to the original OECD report.

Friday April 4, 2014 • 4:20 PM - 6:00 PM, Regency A
Coordinated Session, E2

Improving Technical Quality of Computerized Adaptive Testing in K-12 Assessments

Organizer and Session Chair: Liru Zhang, Delaware State Department of Education, Dover, DE

Modeling Longitudinal Construct of Achievement Computerized Adaptive Tests

Shudong Wang, NWEA, Portland, OR, Liru Zhang, Delaware Department of Education, Dover, DE, Gregg Harris, NWEA, Portland, OR, and Hong Jiao, University of Maryland, College Park, MD

The study uses the multiple-indicator, latent-growth modeling (MLGM) approach to examine the longitudinal achievement construct and its invariance for a large scale standardized computerized adaptive mathematics test in K-12 education. The results of the analyses suggest that with repeated measures, the construct of test remained consistent at different time points.

Investigating Comparability of Computerized Adaptive Tests

Liru Zhang, Delaware Department of Education, Dover, DE, Allen Lau, Pearson, San Antonio, TX, and Shudong Wang, NWEA, Portland, OR

This study explores the comparability of reading scores derived from an adaptive test in a high-stakes state assessment. Evidence is collected through empirical analyses and professional judgment to evaluate the similarity in test construct across individual tests at various ability levels and their effects on the comparability of test scores.

Identifying Item Compromisation in CAT for a K-12 Assessment

Jinming Zhang, University of Illinois at Urbana-Champaign, Champaign, IL

A sequential procedure based on IRT is developed to identify compromised items in a CAT system. It is then applied to real data from a K-12 CAT system, where the order of the students taking the test may be by class or by school, and thus may not be random.

Investigating Optimal Item Selection Procedures for Mixed-Format CAT

Yi Du, Data Recognition Corporation, Shoreview

This study is to compare item selection procedures, i.e., Fisher's information, Kullback-Leibler information, and global information under Bayesian concepts across ability estimation methods, MLE, WLE, and EAP, in a mixed-format CAT to clarify discrepancies in the implementations of different procedures and understand mixed-format CAT in a K-12 measurement setting.

Friday April 4, 2014 • 4:20 PM - 6:00 PM, Regency B
Paper Session, E3

Some Big Ideas in Innovative Measures: Ubiquitous Assessment, National Norms, Learning Progressions, Networked Reading Comprehension

Session Chair: Weiwei Cui, NISS, Research Triangle Park, NC

Consequential Decisions From Continuously-Gathered Electronic Interactions: Could it Really Work?

Randy Bennett, ETS, Princeton, NJ

Continuously embedded electronic (or "stealth") assessment has been heralded as an innovation likely to significantly disrupt educational measurement, even leading to the demise of testing as we know it. For formative purposes, such an approach has much to offer. But for summative purposes, could it really work?

Developing National Norms From an Existing Database of Test Scores

Joshua Tudor and Stephen B. Dunbar, The University of Iowa, Iowa City, IA,

This research concerns developing national achievement test norms using population data that are unrepresentative of the nation. Weighted results from three sampling plans differing in the unit of sampling are evaluated against established national norms. The importance of norms and the efficacy of estimating them from existing data are discussed.

An Empirical Approach to Developing Learning Progressions

Philip Giesy, Renaissance Learning, Inc., Vancouver and Laurie Borkon, Renaissance Learning, Madison, WI

The paper discusses new methodologies for analyzing and improving learning progressions, which employ a combined qualitative/quantitative evaluation process. Qualitative input comes from pedagogically informed placement of skills in a learning progression. Quantitative input comes from an extensive database of psychometrically derived item difficulty values for the same skills.

Psychometric Invariance of Online Reading Comprehension Assessment Across Measurement Conditions

Weiwei Cui and Nell Sedransk, NISS, Research Triangle Park, NC

Online reading comprehension assessments (ORCA) were administered under three different measurement conditions: real internet access, restricted internet access, and no internet access. Multidimensional IRT modeling was used to investigate the psychometric invariance of ORCA at model and item levels when identical or equivalent tests were administered in different measurement conditions.

Friday, April, 4, 2014 • 4:20 PM - 6:00 PM, Regency C1
Paper Session, E4

Mobile Devices: Results of Assessment on Tablets

Session Chair: James B. Olsen, Renaissance Learning, Orem, UT

Score Comparability for Web and iPad Delivered Adaptive Tests

James B. Olsen, Renaissance Learning, Orem, UT

Score comparability was evaluated in parallel studies of web and iPad applications using common item banks, adaptive algorithms and scoring. Rasch Score correlations across applications and grades are: Reading 0.936, Math 0.916, Early Literacy 0.757. Scatter plots, paired t-tests, and effect sizes showed score adjustments are not warranted between applications.

Testing on Tablets

Lei Yu, William Lorie, and Les Sewall, Questar Assessment Inc., Apply Valley

To investigate the feasibility and usability of new devices as additional options for high-stakes online assessments, cognitive labs (cog labs) were conducted for a state's high school testing program using iPads and Androids. The findings have important implications for successful deployment of high stakes testing on new devices.

Assessing Students on iPads: A Usability Study

Laura Stoddart, DRC, Maple Grove, MN, Huiqin Hu, DRC, Plymouth, MN, and Jennifer Norlin-Weaver, DRC, Maple Grove, MN

With the increased use of iPads in classrooms, efforts are being made to develop tests that can be delivered on iPads. This usability study investigates whether the iPad-based test assesses the relevant construct. The findings will significantly contribute to the iPad-based test development practice and related decision making policy.

Assessing Student Writing on Tablets

Laurie L. Davis, Aline Orr, Xiaojing Kong, and Chow-Hong Lin, Pearson, Austin, TX

The educational assessment landscape is being transformed by the introduction of tablets. The current study evaluates differences in student writing on tablets vs. laptops and determines to what degree external keyboards for tablets can overcome these differences. Data were collected from 848 students in 5th grade and high school.

Friday, April, 4, 2014 • 4:20 PM - 6:00 PM, Regency C2
Coordinated Session, E5

How Comparable are International Comparisons of Educational Outcomes?

Organizer and Session Chair: Wolfram Schulz, ACER, Camberwell, Victoria, Australia

Measurement Equivalence of Attitudinal Items: Evidence From ICCS 2009

Maria Magdalena Isac, CRELL, Centre for Research on Lifelong Learning, Ispra, Italy and Dorota Weziak-Bialowolska, European Commission JRC, Ispra, Italy

In this study, we argue that cross-country comparisons of attitudinal measures in cross-national studies are appropriate only if the instruments meet the measurement equivalence requirements. Within a multi-group factor analytic framework, we illustrate with the citizenship values scales from ICCS 2009 possibilities and limitations in achieving meaningful cross-country comparisons.

Age-Based and Grade-Based Sampling in International Surveys on Education

Wolfram Schulz, ACER, Camberwell, Victoria, Australia

International studies such as PISA and TIMSS aim to compare students' learning outcomes across a variety of national education systems and over time. This paper illustrates differences in sampling methodology across these studies and discusses the implications of these differences for comparisons and secondary data analysis.

Alternative Ways of Measuring Attitudes in Cross-National Studies: Examples From PISA 2012

Petra Lietz, Australian Council for Educational Research, Adelaide, South Australia, Australia and Jonas Bertling, ETS, Princeton, NJ

Differences in response style to Likert-type items across countries is a known phenomenon. Still, the extent to which alternative item formats can address this issue has not been resolved. Using field trial and main survey data from PISA 2012, analyses are undertaken to address this issue for attitudes towards mathematics.

Measuring Cognitive Outcomes Across Diverse Educational Contexts

Julian Fraillon and Tim Friedman, Australian Council for Educational Research, Melbourne, Australia

ICCS and ICILS assess outcomes in cross-curricular non-traditional learning areas that are rapidly changing. This presentation describes how the processes of assessment framework development, test construction and review of item data can contribute to the establishment of cross-nationally equivalent measures of learning outcomes in these areas of international assessment.

Friday April 4, 2014 • 4:20 PM - 6:00 PM, Commonwealth A Coordinated Session, E6

Modern Profile Analysis via Multivariate Statistics (PAMS)

Organizer and Session Chair: Se-Kang Kim, Fordham University, Bronx, NY

Predictive Validity: When are Subscores Important?

Mark L. Davison and Ernest C. Davenport, University of Minnesota, Minneapolis, MN

This paper describes regression procedures for assessing whether subscores contribute to the prediction of a criterion variable over and above the contribution of the total score. It is concluded that subscores contribute to predictive validity when there is a subscore pattern that distinguishes high and low scores on the criterion.

Longitudinal English Language Learns' Language Proficiency Using the PAMS Model

Daeryong Seo, Husein Taherbhai, Pearson, San Antonio, TX; and Se-Kang Kim, Fordham, Bronx, NY

PAMS was applied into longitudinal English Language Proficiency (ELP) subtest scores. The results indicated that the first profile was a negative indication for ELA achievement but the second profile of yearly growth of Speaking plus Earlier Reading and Later Writing was a positive indication of ELA achievement.

Multidimensional Profile Pattern Analysis of Noise in Cognitive Diagnostic Models

Joe Grochowalski and Se-Kang Kim, Fordham University, Bronx, NY

Using Canonical Correspondence Analysis, we examine the multidimensional response profiles that are unexplained by the Q-matrix from Tatsuoka (1990) Fraction Subtraction data. Rather than modeling (and assuming) slipping and guessing as noise, these descriptive residual profiles provide insight into the source of the unexpected responses including Q-matrix misspecification.

Growth Pattern vs. Growth Level: What Do They Tell

Cody Ding, University of Missouri at St. Louis, St. Louis, MO

The purpose of the paper is to study and compare growth trajectories with respect to pattern and level in latent growth modeling via structural equation modeling (SEM) and multidimensional scaling (MDS) based approaches. These two approaches are discussed using a data on children's reading growth.

Friday April 4, 2014 • 4:20 PM - 6:00 PM, Commonwealth B Coordinated Session, E7

Extensions to Evidence Based Standard Setting

Organizer: Laurie L. Davis, Pearson, Austin, TX
Session Chair: Natasha J. Williams, Pearson, Austin, TX

An Example of EBSS for an English Language Proficiency Assessment

Sonya Powers, Natasha J. Williams, Leslie Keng, and Laura Starr, Pearson, Austin, TX

After transitioning to a new academic assessment, Texas began reviewing the standards on their English language proficiency test to evaluate the alignment of performance standards on the two assessments. This paper describes the content and empirical analyses used during the evidence based standard setting process.

Standard Setting for a Common Core Aligned Assessment

Ye Tong, Pearson, Audubon, PA, Brian Patterson, College Board, New York, NY, Peter Swerdzewski, Regents Research Fund, New York, NY, and Candy Shyer, NYSED, Albany, NY

The New York State grades 3–8 assessment transitioned to the common core in the spring of 2013. An evidence-based standard setting process was carried out where empirical benchmark studies were incorporated to help define college readiness, along with a well-research standard setting methodology.

Lessons Learned: Decision Points for Empirical-Based Standard Setting

Leslie Keng, Natasha J. Williams, Pearson, Austin, TX; and Sonya Powers, Pearson, Iowa City, IA

This paper discusses decision points that testing programs face as they implement the empirical-based standard-setting (EBSS) approach to establish standards for their assessments. The three different implementations of EBSS presented in this session and implementations previously presented or published are synthesized and compared in several key areas.

Evidence-Based Standard Setting: Vertically Aligning Grades 3–8 Assessments

Aimee Boyd, Laurie Davis, Sonya Powers, Robert Schwartz, and Ha Phan, Pearson, Austin, TX

The evidence-based standard setting process is extended to the Texas assessments in grades 3–8 resulting in a comprehensive assessment system with vertically aligned performance standards that anchor at high school and link back to grades 3–8. Policy considerations, educator committees, and empirical studies are integral throughout the process.

Friday, April, 4, 2014 • 4:20 PM - 6:00 PM, Commonwealth C
Paper Session, E8

Multilevel Models

Session Chair: Ji Seung Young, University of Maryland, College Park, MD

Multilevel Modeling of Cognitive Diagnostic Assessment

Xue-Lan Qiu and Wen-Chung Wang, Hong Kong Institute of Education, Hong Kong, Hong Kong

Most cognitive diagnostic models are unilevel. This study develops a multilevel version of the log-linear cognitive diagnosis model. The simulation results showed that the parameters and latent attribute profiles were recovered fairly well; and ignoring multilevel structure would underestimate the standard error and yield a poor classification rate.

Handling Correlations Between Covariates and Random Slopes in HLMs

Michael Bates, Michigan State University, East Lansing, MI, Katherine Furgol Castellano, Educational Testing Service, San Francisco, CA, Sophia Rabe-Hesketh, University of California at Berkeley, Berkeley, CA, and Anders Skrondal, Norwegian Institute of Public Health, Oslo, Norway

It is well-known that correlations between covariates and random intercepts in hierarchical linear models (HLMs) can lead to bias, but the problem of correlations between covariates and random slopes is rarely considered. We investigate the bias of standard estimators in this case and propose a consistent two-stage estimator.

Incorporating Mobility for Multilevel and Repeated Item Response Data

In-Hee Choi, University of California at Berkeley, Graduate School of Education, Albany, CA and Mark Wilson, University of California at Berkeley, Graduate School of Education, Berkeley, CA

The primary goal of this study is to propose cross-classified multiple membership models to analyzing longitudinal item response data in which students switch schools between measurement occasions. Furthermore, this study investigates the impacts of misspecifications of school membership in the analysis of longitudinal data sets that include mobile students.

Estimation of a Three-Level Latent Model With Metropolis-Hastings Robbins-Monro Algorithm

Ji Seung Yang, University of Maryland, College Park, MD and Li Cai, UCLA, Los Angeles, CA

A Metropolis-Hastings Robbins-Monro algorithm (MH-RM; Cai, 2008, 2010a, 2010b) is implemented to estimate cross-level interactions in a three-level nonlinear latent variable model with more computational efficiency. Both one- and two-stage estimation methods are considered, and preliminary results show that the algorithm can obtain full information maximum likelihood (FILM) estimates properly.

Friday, April, 4, 2014 • 4:20 PM - 6:00 PM, Commonwealth D
Paper Session, E9

Solutions for Difficult Problems and Unusual Data

Session Chair: Dan McNeish, University of Maryland, College Park, MD

Bayesian Estimation of Random Coefficients in Polytomous Response Models

Hariharan Swaminathan and Jane Rogers, University of Connecticut, Storrs, CT

In multilevel regression models, the first level random regression coefficients are critical for prediction and validation of tests. Pure Bayesian estimation of these coefficients for polytomous responses is provided and illustrated with assessing (a) factors influencing DIF,(b) accountability and instructional sensitivity, and (c) prediction of future student proficiency categories.

A Class of Unfolding Models for Pairwise-Comparison and Ranking Data

Chen-Wei Liu and Wen-Chung Wang, The Hong Kong Institute of Education, Hong Kong, Hong Kong

Pairwise comparison and ranking data are commonly used. Unfolding IRT models may be more appropriate for attitude items than dominance ones. We thus developed a class of IRT unfolding models for pairwise comparison and ranking data. The simulation results showed the parameters could be recovered fairly well.

Estimating Item Parameters for Small Samples With Kaplan-Meier

Dan McNeish, University of Maryland, College Park, MD

Kaplan-Meier is a popular nonparametric method for analyzing time-to-event data. It will be argued that item-response data can be viewed as time-to-event data so that Kaplan-Meier can be applied. Producing item parameters with Kaplan-Meier will be discussed and simulation studies are provided to compare these estimates to Bayesian MCMC estimates.

Estimating 2PL IRT Parameters From 3PL Ones Without Response Data

Zhiming Yang, Educational Testing Service, Newtown, PA; Lin Wang, and Shelby J. Haberman, Educational Testing Service, Princeton, NJ

A logistic regression method and a simulation method were used to estimate 2PL IRT item parameters from the items' existing 3PL parameter estimates when response data were not available. The two methods yielded similar 2PL item parameter estimates and raw-to-scaled score conversion tables.

Friday April 4, 2014 • 6:30 PM-8:00 PM, Howe Room
33rd Floor, Loews Hotel

NCME and AERA Division D Joint Reception

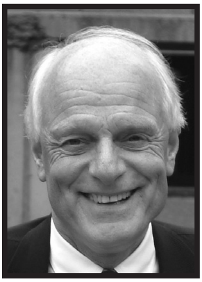
Saturday, April 5, 2014 • 7:45AM-9:00 AM
Regency Ballroom B

2014 NCME Breakfast and Business Meeting

Join your friends and colleagues at the NCME Breakfast and Business Meeting at the Loews 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 5, 2014 • 9:00 AM-9:40 AM
Regency Ballroom B

Presidential Address



Scaling Test Items: Top Down or Bottom Up?

Wim van der Linden
CTB/McGraw-Hill

Saturday, April 5, 2014 • 10:00 AM-11:40 AM, Commonwealth A
Invited Session, F1

NCME Book Series: New NCME Applications of Educational Measurement and Assessment Book Series

Organizer: Michael Kolen, University of Iowa
Session Chair: Wayne Camara, The College Board

The purpose of this session is to introduce the NCME membership to the new NCME Book Series, which is intended to increase the understanding and dissemination of research-based applied educational measurement and assessment. In this session, the Book Series is described followed by a discussion of three volumes that are under development.

Introduction to Session and Description of the Book Series
Wayne Camara, The College Board and Michael Kolen, University of Iowa

Wayne Camara introduces the session and provides a brief overview of the history of the Book Series, which was initiated when he was NCME President. Michael Kolen, Editor of the Book Series, describes the book series, including characteristics of volume editors and how ideas for new volumes are initiated.

Meeting the Challenges to Measurement in an Era of Accountability
Henry Braun, Boston College

Henry Braun describes the volume that he is editing, titled "Meeting the Challenges to Measurement in an Era of Accountability." This volume is scheduled to be published on January 1, 2015.

Technology and Testing: Improving Educational and Psychological Measurement
Fritz Drasgow, University of Illinois, Urbana-Champaign

Fritz Drasgow describes the volume that he is editing, titled "Technology and Testing: Improving Educational and Psychological Measurement." This volume is scheduled to be published on October 1, 2014.

Fairness
Neil Dorans, Educational Testing Service

Neil Dorans describes the volume that he and Linda Cook are editing, tentatively titled "Fairness." This volume is tentatively scheduled to be published on April 1, 2016.

Saturday, April 5, 2014 • 10:00 AM-11:40 AM, Commonwealth D
Award Winning Research, F2

Award Winning Research

Organizer and Session Chair: Jill van den Heuvel, Alpine Testing Solutions, Inc.

Reporting Reliable Change in Students' Overall and Domain Abilities Across Two Time Points
Chun Wang, University of Minnesota

Measurement of change in student performance is pivotal in educational research. Average growth differs from one sub-content area to another. This study presents a longitudinal extension of a higher-order IRT model, and shows how this new approach could improve the reliability of change score at both overall and domain levels.

Sharing Research and Beyond
Kyung (Chris) T. Han, Graduate Management Admission Council

Kyung T. Han, a winner of the Jason Millman Promising Measurement Scholar Award, will share his story about his research and development of software tools. Several psychometric software tools that he developed, including WinGen, IRTEQ, SimulCAT, MSTGen, and SimulMAT, will be introduced as well as his latest work on CAT.

Optimized Item Pool Generation and the Performances of Multidimensional CAT
Lihua Yao, Defense Manpower Data Center

Optimized item pools are generated by varying the sample sizes with the expected information. The relation between the pool size, sample size, and test length, with or without item exposure control, are studied. The generated item pools are compared with the real pool through the performances of MCAT and UCAT.

Investigation of Optimal Design and Scoring for Adaptive Multistage Testing (MST): A Tree-Based Regression Approach
Duanli Yan, Educational Testing Service

The dissertation introduced a new nonparametric tree-based MST methodology. This new algorithm has advantages over the traditional methodologies to MST, including simplicity, lack of restrictive assumptions, and the possible implementations based on small samples. It is an extension of the tree-based CAT design described by Yan, Lewis and Stocking (2004).

Saturday, April 5, 2014 • 10:00 AM - 11:40 AM, Regency C1
Coordinated Session, F3

Evaluating the Evaluations: Exploring Improvements to Measuring Classroom Mathematics Instruction

Organizer and Session Chair: Mark Chin, Harvard University, Cambridge, MA

Using Surveys as Proxies for Observations in Measuring Mathematics Instruction

David Braslow, Harvard Graduate School of Education, Cambridge, MA and Andrea Humez, Boston College Lynch School of Education, Newton, MA

Using data from elementary mathematics teachers, we examine the correspondence between self-reports and observational measures of two instructional dimensions—reform-orientation and classroom climate—and the relative ability of these measures to predict teachers' contributions to student learning.

Dimensionality and Generalizability of the Mathematical Quality of Instruction Instrument

Ben Kelcey, University of Cincinnati, Cincinnati, OH, Dan McGinn, Harvard University, Cambridge, MA, Heather Hill, Harvard Graduate School of Education, Cambridge, MA, and Charalambos Charalambous, University of Cyprus, Nicosia, Cyprus

The purpose of this study was to investigate three aspects of construct validity for the Mathematical Quality of Instruction classroom observation instrument: (1) the dimensionality of scores, (2) the generalizability of these scores across districts, and (3) the predictive validity of these scores in terms of student achievement.

Impacts on Evaluations: Exploring Rater Ability Identifying Effective Mathematics Instruction

Mark Chin, Cynthia Pollard, Harvard University, Cambridge, MA, Mary Beisiegel, Oregon State University, Corvallis, OR, and Heather Hill, Harvard Graduate School of Education, Cambridge, MA

While considerable variance in teachers' scores on observational instruments is attributed to raters, rater accuracy and its impact on score quality remains underexplored. Using student achievement data and ratings of mathematics instruction, we study methods for differentiating raters by accuracy and investigate whether these differences affect reliability and validity.

Using Item Response Theory to Learn About Observational Instruments

Dan McGinn, Harvard University, Cambridge, MA, Ben Kelcey, University of Cincinnati, Cincinnati, OH, Heather Hill, Harvard Graduate School of Education, Cambridge, MA, and Mark Chin, Harvard University, Cambridge, MA

As many states are slated to soon use scores derived from classroom observation instruments in high-stakes decisions, developers must cultivate methods for improving the functioning of these instruments. We show how multidimensional, multilevel item response theory models can yield information critical for improving the performance of observational instruments.

Saturday, April 5, 2014 • 10:00 AM - 11:40 AM, Regency C2
Paper Session, F4

Item and Rater Drift

Session Chair: Jodi M. Casabianca, University of Texas at Austin, Austin, TX

Detecting Drifted Polytomous Items: Using Global Versus Step Difficulty Parameters

Xi Wang, University of Massachusetts at Amherst, Sunderland, MA and Ronald K. Hambleton, University of Massachusetts at Amherst, Amherst, MA

This study is aimed at investigating the performance of four methods in detecting item parameter drift in polytomous items. The four methods include using item step difficulty thresholds, global difficulty parameter, global difficulty weighted by the score points, and using the difference between item characteristic curves.

A Stepwise Test Characteristic Curve Method to Detect Item Parameter Drift

Rui Guo, Yi Zheng, and Huahua Chang, University of Illinois at Urbana-Champaign, Champaign, IL

This paper introduces a method of item parameter drift detection that can dynamically flag drifted items. The simulation study results showed that this method can perform well in paper and pencil based test.

Analysis of Potential Causes of Item Parameter Drift

Ethan Arenson, Haiyan Wu, and Hao Song, National Board of Osteopathic Medical Examiners, Chicago, IL

Item parameter drift (IPD) occurs when item parameter estimates from subsequent calibrations substantially deviate from the estimates from their first calibration. Severe IPD can adversely impact the estimation of ability scores. This paper attempts to identify the relationship between IPD and item features for a licensure examination.

Augmented Generalizability Study Models for Analyzing Rater Drift

Jodi M. Casabianca, University of Texas at Austin, Austin, TX; J.R. Lockwood, and Daniel F. McCaffrey, Educational Testing Service, Princeton, NJ

This research uses augmented generalizability study models that include piecewise-polynomial "B-spline" basis functions to estimate the amount of variance from time trends. We apply the model to scores from classroom observation protocols administered in the Understanding Teaching Quality study to investigate and understand the nature of rater changes over time.

Saturday April 5, 2014 • 10:00 AM - 11:40 AM, Washington A
Paper Session, F5

Goodness of Fit Statistics / Propensity Score Matching

Session Chair: Peter van Rijn, ETS Global, Amsterdam, Netherlands

Assessing Item Fit Using Residuals with Applications to NAEP

Sandip Sinharay, CTB/McGraw-Hill, Monterey, CA, Shelby Haberman, ETS, Princeton, NJ, and Kyong-Hee Chon, Western Kentucky University, Bowling Green, KY

We suggest a new form of residual analysis to assess item fit for unidimensional IRT models. The large sample distribution of the residual is proved to be standardized normal when the IRT model fits the data. We apply the residual analysis to data from the National Assessment of Educational Progress (NAEP).

A Comparison of Methods for Assessing Factor Structure in Multidimensional IRT

Richard Schwarz, ETS, Salinas, CA and Lihua Yao, DMDC, Monterey, CA

Multidimensional IRT uses several methods to determine the number and types of dimensions to retain such as global fit or visual ones (e.g., direction cosines). Using simulation and operational data sets, the correspondence between different measures is presented and guidance developed for collectively evaluating different types of multidimensional output.

Pragmatic versus Statistical Strategies for Log-linear Smoothing Model Selection

Yi Cao, Tim Moses, and Andrea Bontya, ETS, Princeton, NJ

This study compares a pragmatic sample size strategy, developed to accommodate operational constraints, with other log-linear smoothing model selection strategies (e.g., likelihood ratio test and AIC), and examines their influence on equating regarding conversions and examinees' pass/fail decisions. Preliminary results support the use of the pragmatic sample size strategy operationally.

Extensions of Generalized Residual Analysis for Assessing IRT Model Fit

Peter van Rijn, ETS Global, Amsterdam, Netherlands, Sandip Sinharay, CTB/McGraw-Hill, Monterey, CA, and Shelby Haberman, Educational Testing Service, Princeton, NJ

We extend the method of generalized residuals for assessing model fit developed by Haberman (2009) to incomplete designs, complex sampling, and multidimensional IRT models. The residuals are approximately standard normal. The method is applied to data from the National Assessment of Educational Progress program. Implications of misfit are discussed.

Saturday, April 5, 2014 • 10:00 AM - 11:40 AM, Commonwealth B
Coordinated Session, F6

Designing System of Next Generation Science Assessments: Challenges, Choices, Trade-Offs

Organizer and Session Chair: Pascal Forgione, ETS, Austin, TX

Inherent Measurement Challenges in NGSS for Formative and Summative Assessments

Joanna Gorin, ETS, Princeton, NJ

NGSS are ambitious in their goals for and demands of teaching and assessment. This presentation examines NGSS assessment as a complex system, with all of its inherent, competing goals, and will discuss the major challenges for building appropriate assessment tools and key strategies offering the greatest promise.

Potential Comprehensive Science Assessment System Designs and Their Trade-Offs

Nancy Doorey, K-12 Center, ETS, Wilmington, DE and Kathleen Scalise, University of Oregon, Eugene, OR

This presentation focuses on "What might a next generation science assessment system look like?", and uses very distinct system designs: matrix sampling of extended hands-on tasks or simulations; a through-course design; or possibly a structure that allows each student to progress at their own pace with only periodic standardized examinations.

Critical Perspectives on NGSS and NGSA: Curriculum and Measurement

Andrew Porter, University of Pennsylvania, Philadelphia, PA and Mark Wilson, University of California at Berkeley, Berkeley, CA

Two experts representing the curriculum/standards and the measurement/assessment fields will offer critical commentary on the new aspects of these standards and the challenges that educators and the measurement community are likely to confront when building instructional and assessment resources, especially the integration across three dimensions of the NGSS.

The Next Generation Science Standards: How Do They Differ?

Pascal D. Forgione, K-12 Center, ETS, Austin, TX and Nancy Doorey, K-12 Center, ETS, Wilmington, DE

NGSS contain new ideas for how and what students should learn in Science. While exciting, new standards also bring with them a host of challenges for educators. This opening presentation will set a critical context for understanding the uniqueness of three dimensional NGSS adopted by some two dozen "lead" states.

Saturday, April 5, 2014 • 10:00 AM - 11:40 AM, Commonwealth C
Coordinated Session with Special Format, F7

What is the Best Way to Use the Term 'Validity'?

Organizer: Paul E. Newton, Institute of Education, University of London (UK)
Session Chair: Linda Cook, Educational Testing Service

Debate over the best way to use the term 'validity' has reignited in recent years. Some theorists insist that validity is essentially a matter of interpretation (measurement); others insist that validity is essentially a matter of use (decision-making). Some believe that the term 'validity' can legitimately be applied to both interpretations and uses; others believe that the very debate over how best to apply the term 'validity' is problematic, if not misguided. This session involves contributions from theorists from across the spectrum of perspectives. It invites NCME colleagues to engage with this debate, to explore possibilities for its reconciliation.

Problems and Pseudo-Problems in Test Validity Theory
Keith A. Markus, John Jay College of Criminal Justice, CUNY

Validation of Score Inferences is Different From Justification of Test Use
Gregory J. Cizek, University of North Carolina, Chapel Hill

Shifting Foci of Validity for Test Use
Pamela A. Moss, University of Michigan, Ann Arbor

Debunking the Debate: Validity Refers to Test Use
Stephen G. Sireci, University of Massachusetts, Amherst

Validating the Interpretation and Use of Scores
Michael T. Kane, Educational Testing Service

Do We Need to Use the Term 'Validity'?
Paul E. Newton and Stuart D. Shaw, Institute of Education, University of London (UK), Cambridge International Examinations (UK)

Reflective Overview
Lorrie A. Shepard, University of Colorado, Boulder

Saturday, April 5, 2014 • 10:00 AM - 11:40 AM, Washington B
Paper Session, F8

Missing Data

Session Chair: Nathaniel J.S. Brown, Boston College, Chestnut Hill, MA

Robust Growth Mixture Models With Non-Ignorable Missingness

Zhenqiu (Laura) Lu, University of Georgia, Athens, GA and Zhiyong Zhang, University of Notre Dame, Notre Dame, IN

Four non-ignorable missingness models to recover the information due to missing data and three robust models to reduce the effect of non-normality were proposed. A Bayesian method was implemented. Simulation studies and real data analysis were conducted to evaluate their performances.

Modeling Missing-Data Processes: An IRTree-Based Approach

Dries J.L. Debeer, Rianne Janssen, University of Leuven, Leuven, Belgium; and Paul De Boeck, Ohio State University, Columbus, OH

In large-scale assessments two missing responses can be discerned: items can be "not-reached" or "skipped". Both omissions may be related to the test takers' proficiency, resulting in non-ignorable missingness. An IRTtree-based model with separate latent processes for "not-reached" and "skipped" items is proposed, and illustrated using empirical data.

Comparison of Likelihood Functions of Multiple Imputation for Nested Models

Yoonsun Jang, Zhenqiu (Laura) Lu, and Allan Cohen, University of Georgia, Athens, GA

Three types of pooling likelihood functions for multiply imputed data are calculated and investigated for hierarchical linear models. In addition, four types of weights are applied to these likelihood functions, and nested models are compared. Simulation studies showed no pooling method or weight uniformly performed well for all conditions.

Impact of Omitted Responses Scoring Methods on Achievement Gaps

Nathaniel J.S. Brown, Boston College, Chestnut Hill, MA; Dubravka Svetina and Shenghai Dai, Indiana University, Bloomington, IN

Several methods exist for dealing with intentionally omitted responses in achievement tests, but little is known about their practical impact when student groups omit responses at different rates. Four methods were applied to the 2009 NAEP Mathematics Assessment to determine their impact on reported achievement gaps.

Saturday, April 5, 2014 • 11:50 AM-12:40 PM, Washington
Invited Session, G1

Session Chair: Matthew Johnson, Teachers College, Columbia University, New York, NY

A Location Scale Item Response Theory (IRT) Model for Ordinal Questionnaire Data

Donald Hedeker
University of Illinois at Chicago

An ordinal IRT model is described which allows within-subjects variance and random subject scale. In addition to item parameters associated with a subject's ability, item difficulty and discrimination scale parameters are included. These scale parameters indicate the degree to which items are scaled differently across the ordinal categories (scale difficulty) and separate subjects of varying levels of variability (scale discrimination).

Saturday, April 5, 2014 • 11:50 AM-12:40 PM, Regency B
Invited Session, G2

Session Chair: Diego Zapata-Rivera, ETS, Princeton, NJ

Model-Based Tools Embedded Within Games to Assess and Support Important Competencies

Valerie Shute
Florida State University

Psychometrics involves the design, administration, and interpretation of tests to measure psychological variables. And while we can't actually "measure" these constructs directly, we can measure proxies, and make inferences back to the targeted constructs. The inferences can then be used to support learning. I will illustrate this approach in the context of a game we recently developed called Newton's Playground.

Overall Electronic Board Session Chairs: Dylan Molenaar, University of Amsterdam, Amsterdam, Netherlands and Goran Lazendic, Australian Curriculum, Assessment and Reporting Authority, Sydney, Australia

Saturday, April 5, 2014 • 12:50 PM - 2:20 PM, Millennium Hall
Electronic Board Paper Session 11

Method and Model Violations and Distortions

Session Chair: Jorge Gonzalez, Pontificia Universidad Catolica de Chile, Santiago, Chile

Electronic Board Presentation #11.1A (Monitor 1)
Exploration of Subgroup Equating Invariance on Elementary Reading Assessments
Feng Chen and Amy K. Clark, University of Kansas, Lawrence, KS

The present study explores equating invariance for gender and poverty subgroups. Both real-data and simulation analyses were performed to evaluate the magnitude of invariance over multiple test forms and grade levels. RMSD values generally fell within expected range based on previous findings.

Electronic Board Presentation #11.2A (Monitor 2)
Impact of the Equating Sample's Mode Composition on RASCH Equating
Shameen N. Gaj, Junhui Liu, and Hyeonjoo Oh, ETS, Princeton, NJ

When moving tests from paper to online administration, what proportion of the equating data should represent each mode? Preliminary results from this study found that the equating data should represent the proportion of each mode in the population. The results will highlight this importance with respect to mode effects.

Electronic Board Presentation #11.1B (Monitor 1)
Comparison of IRT Preequating Methods When Item Positions Change
Yong He and Zhongmin Cui, ACT, Inc., Iowa City, IA

This study investigates which IRT equating method is most robust in the context of item position changes through real data analyses. Test forms were administered in a spiraled manner to control context effects other than position changes. Preequating results were compared with postequating results for evaluation.

Electronic Board Presentation #11.2B (Monitor 2)
Violating the Unidimensional Assumption of Item Response Theory Vertical Scaling
Anna M. Topczewski, Pearson, Ann Arbor, MI

Research has shown UIRT vertical scaling methods can lead to inconsistent results. This research investigates the effect violating the UIRT vertical scaling unidimensionality assumption. Given the simplicity of UIRT, it is argued that error incurred by violating the unidimensionality assumption could be preferred over MIRT vertical scaling methods.

Saturday, April 5, 2014 • 12:50 PM - 2:20 PM, Millennium Hall
Electronic Board Paper Session 12

Item Parameter Approaches

Session Chair: Jeffrey M. Patton, University of Notre Dame, Notre Dame, IN

Electronic Board Presentation #12.1A (Monitor 3)
Modeling Binary Responses of Common Stimulus Items Using MIRID Model
Shudong Wang, NWEA, Portland, OR and Hong Jiao, University of Maryland, College Park, MD

This study uses simulation method to explore the application of model with internal restrictions on item difficulties as a componential model for common stimulus item (CSI) that are required by the Common Core State Standards (CCSS) and used by the Smarter Balanced Assessment Consortium (SBAC).

Electronic Board Presentation #12.2A (Monitor 4)
Modeling Reading Difficulty of Subject-Matter Assessments
Ting Zhang, American Institutes for Research, Greenbelt, MD; Robert Mislevy, and Xueli Xu, Educational Testing Service, Princeton, NJ

The research aims to measure and partial out construct-irrelevant variance associated with reading in large-scale subject-matter assessments. Using TIMSS science and CIVED test items, the study gauged item reading demand through Coh-Metrix and human rating. By modeling reading difficulty through a MIRT model, the quality of domain proficiency estimates improved.

Electronic Board Presentation #12.1B (Monitor 3)
Reducing the Effects of Careless Response Behavior on Item Calibration
Jeffrey M. Patton, Ying Cheng, Ke-Hai Yuan, University of Notre Dame, Notre Dame, IN; and Qi Diao, CTB/McGraw-Hill, Monterey, CA

Studies have demonstrated that careless response behavior may have negative effects on item parameter estimates, but little research has investigated ways to reduce these effects. In this study, we simulate a number of plausible careless response behaviors and evaluate several methods to reduce their effects on item calibration.

Saturday, April 5, 2014 • 12:50 PM - 2:20 PM, Millennium Hall Electronic Board Paper Session 13

Software Comparisons

Session Chair: Yi Du, Data Recognition Corporation, Maple Grove, MN

Electronic Board Presentation #13.1A (Monitor 5)
A Comparison of Commercial Software for Multidimensional IRT Modeling
Kyung (Chris) T. Han, Graduate Management Admission Council, Reston, VA and Insu Paek, Florida State University, Tallahassee, FL

In this study, the authors evaluate three new MIRT software (IRTPRO 2.1 , FlexMIRT , and ESQIRT) with two well-known existing MIRT software (TESTFACT 4 and Mplus 7.11). The programs' performance on the model parameter estimation are compared regarding their built-in estimation algorithms and realistic test conditions via simulations.

Electronic Board Presentation #13.2A (Monitor 6)
A Multiple-Imputation Estimation of Knowledge State in Cognitive Diagnostic Modeling
Wei Tian, National Assessment of Educational Quality, Beijing, China and Tao Xin, School of Psychology, Beijing Normal University, Beijing, China

A new method of knowledge state estimation in cognitive diagnostic models, called multiple-imputation EAP or MAP, fully taking into account of the uncertainty of the item parameter estimates. The results showed that the accuracy of KS estimation depends on the sample size and the magnitude of item parameters.

Electronic Board Presentation #13.1B (Monitor 5)
Item Parameter Recovery Accuracy: Comparing PARSCALE, MULTILOG and flexMIRT
Shuqin Tao, Benjamin Sorenson, Mayuko Simon, and Yi Du, Data Recognition Corporation, Maple Grove, MN

This study attempted to provide procedures (command files and transformation) needed to obtain comparable item parameter estimates from PARSCALE, MULTILOG, and flexMIRT, and compare them on the accuracy of item parameter recovery for various conditions. Results showed that their performance was similar in some conditions but became divergent in others.

Saturday, April 5, 2014 • 12:50 PM - 2:20 PM, Millennium Hall
Electronic Board Paper Session 14

Assessing and Measuring All Students

Session Chair: Edynn Sato, Pearson, San Francisco, CA

Electronic Board Presentation #14.1A (Monitor 7)
Accessibility and Accommodations Online: What Do We Know?
Edynn Sato, Pearson, San Francisco, CA

This paper presents findings from a critical evaluation of research addressing the effects of online accessibility strategies and accommodations on the performance of English learners and students with disabilities. Implications for the education, measurement, and policy communities will be discussed vis-à-vis the current evidentiary foundation available in the literature reviewed.

Electronic Board Presentation #14.2A (Monitor 8)
Construct Validity and Fairness of Technology-Enhanced Items for Visually-Impaired Students
Linette M. McJunkin, Computerized Assessments and Learning, Lawrence, KS; John Poggio and Susan Gillmor, University of Kansas, Lawrence, KS

Advancements in technology have provided a supportive structure for item and assessment development. Technology-enhanced items are becoming common-place to assess student understanding; however, there is uncertainty these novel items are measuring the same construct across different student populations. This study evaluates the validity of technology-enhanced items across three student groups.

Electronic Board Presentation #14.1B (Monitor 7)
Designing Innovative Science Assessments That are Accessible for Blind Students
Eric G. Hansen, Lei Liu, Aaron Rogat, and Mark Hakkinen, ETS, Princeton, NJ

Innovative science assessments must be accessible to individuals who are blind. A team with experience in assessment design, cognitive science, accessibility, and software development, developed a prototype accessible middle school science task, conducted a small usability study with blind students, and developed recommendations for developing accessible innovative science assessments.

Electronic Board Presentation #14.2B (Monitor 8)
Opportunity to Learn and Students with Disabilities' Mathematical Achievement
Stephen N. Elliott, Alexander Kurz, Arizona State University, Tempe, AZ; Gerald Tindal, University of Oregon, Eugene, OR, and Nedim Yel, Arizona State University, Tempe, AZ

Large-scale assessments and accountability systems are predicated on the assumption that all students have the opportunity to learn (OTL) what they are expected to know and tested on. We document the relationship between OTL and 4th-8th graders' academic achievement in mathematics on within-year formative measures and annual state tests.

Saturday, April 5, 2014 • 12:50 PM - 2:20 PM, Millennium Hall
Electronic Board Paper Session 15

Automated Scoring in Multiple Contexts

Session Chair: Syed Muhammad Fahad Latifi, University of Edmonton, AB, Canada

Electronic Board Presentation #15.1A (Monitor 9)
Comparing Human and Automated Essay Scoring for Student With Disabilities
Heather M. Buzick, Maria Elena M. Oliveri, Michael M. Flor, and Yigal Attali, Educational Testing Service, Princeton, NJ

We evaluate human and automated scores of essays for test takers with learning disabilities and ADHD and those without disabilities. We also use an automated annotation system developed with natural language processing methodology to obtain qualitative information about patterns of misspellings and use it to predict human/machine discrepancies.

Electronic Board Presentation #15.2A (Monitor 10)
Processing Evaluation Comments Using Natural Language Processing in Medical Education
Tracey Hillier, Hollis Lai, and Mark Gierl, University of Alberta, Edmonton, Alberta, Canada

Evaluation comments are an important source of feedback in medical education. We propose a method on how teaching evaluations can be evaluated using natural language processing. Using student comments, we demonstrate a three stage process on how evaluation comments can be analyzed, compiled and classified to improve education quality.

Electronic Board Presentation #15.1B (Monitor 9)
Technology Enhanced Scoring of Multilingual Medical Licensing Examination
Syed Muhammad Fahad Latifi, Mark J. Gierl, University of Alberta, Edmonton, AB, Canada; André-Philippe Boulais, and Andre De Champlain, Medical Council of Canada, Ottawa, Ontario, Canada

An open-source natural language processing (NLP) suite was employed to develop item-scoring models for multilingual medical licensing examination. English, French, and machine translated-French responses of constructed-response items were scored automatically. Specific feature extraction and model building strategies resulted in an average human-machine agreement of 96%.

Electronic Board Presentation #15.2B (Monitor 10)
Visualization Techniques for Validation of Automated Scoring Models
Derrick Higgins, Educational Testing Service, Philadelphia, PA

Artificial-intelligence (AI) scoring models must be qualitatively validated to ensure adequate construct representation, but this can be very difficult due to the complexity and dimensionality of the statistical models they use. This paper aims to apply data visualization techniques from computer science to facilitate the inspection of AI models.

Saturday, April 5, 2014 • 12:50 PM - 2:20 PM, Millennium Hall
Electronic Board Paper Session 16

Collaboration and Assessment: Co-Constructing Understanding

Session Chair: Lei Liu, ETS, Princeton, NJ

Electronic Board Presentation #16.1A (Monitor 11)
A Tough Nut to Crack: Measuring Collaborative Problem Solving
Lei Liu, Alina Von Davier, Jiangang Hao, Patrick Kyllonen, and Diego Zapata-Rivera, ETS, Princeton, NJ

This paper presents a conceptual model for collaborative problem solving (CPS) and an online game-like task that applies this model to measure both individuals' cognitive and social skills during the CPS process in the context of science. The measurement is focused on the collaborative discourse and actions in which the team members engage.

Electronic Board Presentation #16.2A (Monitor 12)
Measuring Knowledge in Teams of Interprofessional Health Care Students
Kelly S. Lockeman, Peter Boling, Deborah DiazGranados, Alan Dow, and Moshe Feldman, Virginia Commonwealth University, Richmond, VA

Interprofessional education seeks to improve healthcare by bringing together health professional students to learn and work collaboratively. Currently, most assessment focuses on learner satisfaction and attitudinal changes. We present an approach to measuring collaborative knowledge of students on interprofessional teams that may predict collaborative behaviors important for better patient outcomes.

Electronic Board Presentation #16.1B (Monitor 11)
Peer Effects in the Classroom: Evidence from New Peers
Margarita Pivovarova, University of Toronto, Department of Economics, Toronto, Ontario, Canada

In this paper, I combine an innovative research design and a unique data set to quantify knowledge spillovers (peer effects) among students in a classroom. I find positive, large and statistically significant spillovers from good students on everyone in class. Peer effect is increasing in own ability of a student.

Saturday, April 5, 2014 • 12:50 PM - 2:20 PM, Millennium Hall
Electronic Board Coordinated Session 17

Decision Support and Measurement Insight Through Interactive Statistical Visualization

Organizer and Session Chair: John Behrens, Pearson, Mishwaka, IN

Electronic Board Presentation #17.1A (Monitor 13)
Analytic and Design Principles for Interactive Visualization
John T. Behrens, Pearson, Mishwaka, IN

Whereas new advances in interactive data visualization have gained ground in the general statistical community, corresponding advances in the psychometric community have been less brisk. This presentation aims to close the gap by explaining shifts in thinking regarding statistical computing and analytics and introduce basic issues of human-computer interaction.

Electronic Board Presentation #17.2A (Monitor 14)
Interactive Decision Support for Refining CAT Item Banks
Quinn N. Lathrop, University of Notre Dame, South Bend, IN

CAT item bank refinement requires an iterative process of imposing and relaxing constraints across the multi-dimensional space of content coverage and dimensions of task and design attributes. We present methods of interactive visualization that provide decision support and insight for this multi-stakeholder, multi-dimensional problem.

Electronic Board Presentation #17.1B (Monitor 13)
Test Bank Visualization for User Generated Assessment Environments
Yun Jin Rho, Pearson, Boston, MA

The content of tests provide an opportunity to research the goals and values of the test designers while creating challenges regarding how to analyze test content as data itself. This paper reports on work to analyze such construction patterns by combining corpus processing analytics with data visualization techniques.

Electronic Board Presentation #17.2B (Monitor 14)
Visualization for Personal Standard Setting in a Custom-Made Placement Test
Yuehmei Chien, Pearson, Iowa City, IA

A personal standard setting process is introduced for a custom online adaptive placement testing system. In order to facilitate this process, a visualization tool has been developed experimentally for the instructors to inspect the structure of the custom pool, explore implications of various configurations, and set an appropriate cut score.

Saturday, April 5, 2014 • 12:50 PM - 2:20 PM, Millennium Hall
Electronic Board Coordinated Session 18

Advances in Data Forensic Methodologies

Organizer and Session Chair: Dmitry Belov, Law School Admission Council, Newtown, PA

Electronic Board Presentation #18.1A (Monitor 15)
Identification of Compromised Items in High Stakes Testing
Dmitry Belov, Law School Admission Council, Newtown, PA

Item preknowledge is hard to detect due to multiple unknowns: unknown subsets of compromised items, and unknown groups of examinees (at unknown schools) accessing those items prior to taking the test. This paper demonstrates that disentangling this problem becomes feasible when existing statistical detectors are merged with combinatorial optimization.

Electronic Board Presentation #18.2A (Monitor 16)
Are the Score Gains Suspicious? – A Bayesian Growth Analysis Approach
Xin Lucy Liu, Data Recognition Corporation, Maple Grove, Fu Liu, University of North Carolina - Greensboro, Greensboro, NC, Mayuko Simon, Data Recognition Corporation, Maple Grove, and Zhiyong Zhang, University of Notre Dame, Notre Dame, IN

Using multiple-year longitudinal data from a large-scale standardized test, we propose to fit a Bayesian nonlinear growth model to identify schools with unusual score gains. The posterior predictive distribution of the residuals derived from a school's observed year-to-year performance from the predicted growth curve provides evidence for potential cheating.

Electronic Board Presentation #18.1B (Monitor 15)
Improving the Robustness of Erasure Detection to Scanner Undercounting
James A. Wollack, University of Wisconsin-Madison, Madison, WI, and Dennis Maynes, Caveon Test Security, Midvale

Analyses based on erasure counts are problematic because scanners often undercount the number of erasures. Consequently, statistical models are based upon spurious machine-counted erasure totals, whereas hypotheses are usually tested using hand-counted totals. This paper demonstrates this problem and introduces and evaluates a new measure that is robust to it.

Electronic Board Presentation #18.2B (Monitor 16)
Comparison of Two Statistics to Detect Cheating in Multistage Test
Jaehoon Seol, American Institute of CPAs, Ewing, Seonho Shin, Measured Progress Inc., Dover, NH, and Larissa Smith, National Board of Osteopathic Medical Examiners, Conshohocken

This study compares the performance of the Kullback-Leibler divergence (KLD) and the Hellinger distance (HD) in detecting answer copying in multistage testing. Unlike KLD, HD satisfies all qualitative properties of the similarity relationship. We first investigate statistical properties of HD, and then compare detection rate between KLD and HD.

Saturday, April 5, 2014 • 12:50 PM - 2:20 PM, Millennium Hall
Electronic Board Coordinated Session 19

Cognitive Diagnostic Models to Address Issues in Next Generation Assessments

Organizer and Session Chair: Burcu Kaniskan, Pearson, San Antonio, TX

Electronic Board Presentation #19.1A (Monitor 17)
Diagnostic Assessment: A Person Fit Study
Mary Roberts and Ying Cui, University of Alberta, Edmonton, Canada

A two-stage procedure was used to evaluate the usefulness of person fit in validating the inferences drawn from cognitive diagnostic tests. First, the hierarchy consistency index was used to identify the misfitting student item-score vectors. Second, students' verbal reports were used to examine the actual causes of misfits.

Electronic Board Presentation #19.2A (Monitor 18)
An Integrated Approach Towards the Development of Cognitive Diagnostic Assessment
Changjiang Wang, Pearson, Iowa City, IA, Laine Bradshaw, University of Georgia, Athens, GA, and James Kowpfler, Pearson, Iowa City, IA

This study illustrates a principled test design procedure in developing a cognitive diagnostic assessment for algebra readiness. An integrated approach is employed, where a blend of the attribute hierarchy method and the loglinear cognitive diagnostic model—called the hierarchical diagnostic classification model—is used to guide the test development process.

Electronic Board Presentation #19.1B (Monitor 17)
Developing a Technologically Enhanced Hybrid: Cognitive Diagnostic Stealth Assessment
Man-Wai Chu and Jacqueline P. Leighton, CRAME/University of Alberta, Edmonton, AB, Canada

The objective of the presentation is to discuss the potential hybridization of cognitive diagnostic assessment (CDA) and stealth assessment (SA) to create CDSA. Four issues are discussed using empirical examples and include: (a) cognitive versus student models, (b) task formats, (c) scoring metrics, and (d) feedback reports to students.

Electronic Board Presentation #19.2B (Monitor 18)
Examining Cognitive Diagnostic Information from Parallel Forms: An Empirical Study
Ruhan Circi, University of Colorado at Boulder, Boulder, CO; Michael J. Young and Burcu Kaniskan, Pearson, San Antonio, TX

The goal of this study is to examine the diagnostic information coming from a subset of mathematics items in two parallel forms with three topics. Two cognitive diagnostic models (CDM) are used for three kinds of comparisons, namely, (a) differences by forms, (b) by models, and (c) subject topic.

Saturday, April 5, 2014 • 2:30 PM - 4:10 PM, Washington A
Paper Session, H1

Multiple Attempts Assessment or Innovative Score Reporting

Session Chair: Claire E. Stevenson, Leiden University, Leiden, Netherlands

Considerations on Including Collaborative Problem Solving Tasks in Standardized Assessments

Alina A. von Davier and Jiangang Hao, Educational Testing Service, Princeton, NJ

In this paper a conceptual framework of an assessment for CPS tasks is presented. Data from an online assessment with two parts, a traditional multiple-choice items part and a game-like CPS task, are discussed and modeled. The data are collected using crowdsourcing via the amazonturk channel.

Examining Analogy Learning as it Occurs With Item Response Theory

Claire E. Stevenson, Leiden University, Psychology Methods & Statistics, Leiden, Netherlands, and Paul De Boeck, Ohio State University, Columbus, OH

Explanatory item response theory models were used to assess children's change during training in analogical reasoning. The utilized IRT models take revised responses after feedback into account. Furthermore, the effectiveness of different types of feedback (metacognitive, cognitive and scaffolds) was evaluated. This paper presents the fitted models and empirical results.

Multiple Attempt Item Functioning in MOOC Formative and Summative Assessments

Kimberly F. Colvin, Massachusetts Institute of Technology, Arlington, MA, Yoav B. Bergner, ETS, Princeton, NJ, and David E. Pritchard, Massachusetts Institute of Technology, Cambridge, MA,

Allowing an examinee to revise an incorrect response improves test information and affords a more reliable ability estimate than only using the examinee's first response (Attali, 2010). These benefits of multiple attempts analysis were compared for both formative and summative assessments in a Massive Open Online Course (MOOC).

Development and Validation of a Technology-Enhanced Score Report

Daniel Lewis, CTB/McGraw Hill, CTB/McGraw Hill, Monterey, CA

The author investigated the use of multimedia to support the interpretation of test results by embedding assessment literacy in a prototype technology-enhanced score. The presenter will demonstrate a technology-enhanced report and discuss the results of a pilot study in which parents viewed technology-enhanced reports during parent-teacher conferences.

Saturday, April 5, 2014 • 2:30 PM - 4:10 PM, Washington B
Paper Session, H2

Comparison Studies

Session Chair: Tianli Li, ACT, Inc., Iowa City, IA

Investigating Scale Comparability Using IRT and Classical Scoring Methods

Tianli Li and Xin Li, ACT Inc., Iowa City, IA

This study examines the scale score comparability using IRT and classical scoring methods for a large scale test over three years, considering the use of various IRT models, IRT theta versus number correct scoring, different IRT and classical equating methods, and different base forms, across multiple subject areas.

Ability Estimation Between IRT Models and Cognitive Diagnostic Models

Yu-Lan Su, ACT, Iowa City, IA and Won-Chan Lee, The University of Iowa, Iowa City, IA

This study analyzes the consistency of ability estimation between the item response theory (IRT) models and cognitive diagnostic models (CDMs). Specifically, the study examines the relationships between examinee expected scores and ability estimates based on IRT models and the mastery level of attributes based on CDMs.

Evaluating Performance Task Scoring Comparability in an International Testing Program

Doris Zahner, CAE, New York, NY and Jeffrey Steedle, Pearson, Austin, TX

The OECD launched AHELO in an effort to measure learning in international postsecondary education. This paper presents a study of scoring equivalence across nine countries for two translated and adapted performance tasks. Results reveal that scorers had similar notions of relative response quality, but not absolute quality.

Investigating Psychometric Isomorphism for a Traditional and Performance-Based Assessment

Derek M. Fay, Roy Levy, Arizona State University, Tempe, AZ; Vandhana Mehta, Independent Researcher, Plano, TX; Allan Reid, Cisco Systems, Inc., Toronto, Canada; David Angus, Unicon, Inc., Gilbert, AZ; Dennis Frezzo, Cisco Systems, Inc., San Francisco, CA; Martin C. Benson, Aries Technology, Inc., Tempe, AZ; and Telethia Willis, Cisco Systems, Inc., Orlando, FL

The psychometric properties for structurally similar tasks potentially differ in meaningful ways. This work pursues several measures that assess key hypotheses about the psychometric comparability of structurally similar tasks for a traditional and performance-based assessment. The relative merits of the proposed measures for different assessment contexts are discussed.

Saturday, April 5, 2014 • 2:30 PM - 4:10 PM, Regency A
Paper Session, H3

Grand Bargain of CAT: Lessons Learned, Field Testing, Challenges & Pool Design

Session Chair: David Shin, Pearson, Iowa City, IA

Applying Computer Adaptive Tests to K–12 Assessments: Advantages and Challenges

Xia Mao and David Shin, Pearson, Iowa City, IA

This study provides some empirical evidence and discussion on the advantages and challenges of transitioning linear computer-based tests (CBT) to computer adaptive tests (CAT) in the context of K–12 assessments using the results of a CAT pilot study from a large-scale state testing program.

Twenty Years of High-Stakes CAT: Lessons Learned in Licensure Testing Can Be Applied to Education

Anthony Zara, Pearson VUE, Bloomington, MN, Phil Dickison, NCSBN, Chicago, IL, Gage Kingsbury, Consultant, Portland, OR, and Jerry Gorham, Pearson VUE, Chicago, IL

The national nurse-licensing program (NCLEX) was one of the first CAT programs, implemented ~20 years ago. This paper reviews NCLEX research findings and provides insight to running a high-stakes CAT program. The lessons learned in licensure testing will be used to provide guidance for the upcoming high-stakes educational accountability tests.

Field Testing in CAT: Sample Size, Location, and Adaptivity

David Shin and Yuehmei Chien, Pearson, Iowa City, IA

This study focuses on the practical field testing issues of the large scale testing program using the empirical CAT settings from a large scale testing program. The study investigates the following factors on FT item calibration quality: (1) sample sizes, (2) FT item insertion locations, (3) administering FT items adaptively.

Item Pool Design for CAT—Review, Demonstration, and Future Prospects

Wei He, NWEA, Portland, OR and Qi Diao, CTB, Monterey, CA

This study will review the literature regarding item pool design for computerized adaptive tests (CATs). A demonstration of the design process to construct the item pools for an operational CAT will be presented by applying the two major methods. Practical Implications and future research directions will be discussed.

Saturday, April 5, 2014 • 2:30 PM - 4:10 PM, Regency B
Coordinated Session, H4

Investigating Genre for Writing Measurement and Automated Writing Evaluation

Organizer: Jill Burstein, ETS, Princeton, NJ
Session Chair: David Williamson, Educational Testing Service, Princeton, NJ

Disjuncture in School and Workplace Writing: A Study of Genre
Jill Burstein, ETS, Princeton, NJ and Norbert Elliot, NJIT, Newark, NJ

Using a survey instrument, presenters explored genre exposure to, and comfort and competency with school and workplace writing. Transactional writing proved to be under-represented in K-12; college students exhibited less comfort and competency with genres such as proposal writing. Expansion of automated systems to evaluate transactional writing genres is proposed.

Does Genre Affect Automated Essay Scoring Performance?
Mark Shermis, University of Akron, Akron, OH

This study used data collected from the Hewlett Prize on automated essay scoring to evaluate structural differences in writing genre for a sample of U.S. high-stakes essay responses. Data were collected from eight tasks reflecting genre differences. Confirmatory factor analysis showed significant differences in the outcomes associated with different genre types.

Workplace Readiness: Aligning Genres to Local and National Expectations
Nancy Coppola, NJIT, Newark, NJ

This presentation focuses on genre and persona in the development of the Technical Communication Body of Knowledge (TCBOK), an international effort sponsored by the Society for Technical Communication (STC). As a designer, the presenter will review the origin and development of validation processes associated with assessment of technical communication.

Genre, Transfer, Distributed Learning: Models for Human and Machine Assessment
Carl Whithaus, UC, Davis, Davis, CA

This paper examines the need to develop models for assessing writing incorporating the multi-genre and multimodal tasks that writers encounter in advanced academic and workplace environments. Within defined validation frameworks, these assessment models can be embedded in calibrated peer review, automated writing evaluation, and hybrid environments.

Saturday, April 5, 2014 • 2:30 PM - 4:10 PM, Regency C1
Coordinated Session, H5

Using Process Data to Assess Student Writing

Organizer and Session Chair: Paul Deane, ETS, Princeton, NJ

Evaluating Writing Process Using Keystroke Logs and NLP Technologies

Gary Feng, Educational Testing Service, Princeton, NJ

We report an approach to evaluate the writing process based on the analysis of individual keystrokes collected in large-scale writing assessments. Natural language processing (NLP) and Bayesian techniques are used to estimate writing processes. We will explore the relation between process variables and scores based on the written product.

Using Writing Process and Product Features to Assess Writing Proficiency

Mo Zhang, Paul Deane, and Yi-Hsuan Lee, Educational Testing Service, Princeton, NJ

In this presentation, we investigated the use of both writing product (that is, the writing piece itself) and writing process features (such as the duration of writing, pauses, and editing) extracted from automated scoring system for assessing test-takers' writing proficiency. Findings based on a variety of analyses will be presented.

Hierarchical Mixture Models for Pause Events During Essay Writing

Russell Almond and Tingxuan Li, Florida State University, Tallahassee, FL

Although writing instruction focuses on writing process, most writing measures focus on the written product (e.g., the essay). Keystroke logs provide information about students' focus during the writing process. This paper fits hierarchical mixture models to pause events and explores associations between model parameters and other measures of writing.

Keystrokes and the Writing Construct: Issues for Analysis

Paul D. Deane, Educational Testing Service, Princeton, NJ

While keystroke logs provide a very detailed window into the actions performed by a writer during the writing process, they are nonetheless indirect forms of evidence about internal cognitive states and processes. We explore some of the key issues that must be addressed.

Saturday, April 5, 2014 • 2:30 PM - 4:10 PM, Regency C2
Paper Session H6

Pre-Equating and Equivalent Group Equating

Session Chair: Xiaoyu Qian, ETS, Princeton, NJ

Section Pre-Equating Under the Equivalent Groups Design Without IRT

Hongwen Guo and Gautam Puhan, ETS, Princeton, NJ

In this paper, we use the randomly equivalent groups design to equate sections of a new test X to an existing test Y. An adjustment is made to obtained the equated scores on the complete form of X based on equated section scores of X, to avoid systematic bias on heuristic approaches some testing programs use in practice.

NEAT and Equivalent Group Equating When Prerequisites are Restricted

Xiaoyu Qian, Jing Miao, and Wenmin Zhang, ETS, Princeton, NJ

The study compared the accuracy of NEAT and equivalent groups (EQ) equating when the number of anchors is restricted and equating groups are close-to-equivalent. The results suggested that NEAT gives consistently more accurate results even with restricted number of anchors. EG equating can tolerate group nonequivalence to a small degree.

Pre-Selecting a Sample for Equivalent Groups in a Nested Design

Ourania Rotou, Xiaoyu Qian, and Chunyi Ruan, Educational Testing Service, Princeton, NJ

Simulated and real data were used to investigate six factors that contribute to the pre-selection of a sample for an equivalent group equating design when examinees are nested in schools. Results suggested that different combinations of the factors of interest may result in obtaining equivalent groups.

Chained Equipercentile Equating With Different Anchor Types

Yanlin Jiang, Haiwen Chen, and Alina von Davier, ETS, Princeton, NJ

The chained equipercentile method can utilize internal, external, or combined anchors to perform equating. This study is to investigate the performance of the CE method using three anchor types under the Kernel equating (KE) framework. The equating results are evaluated in terms of SEEs and RMSE.

Saturday, April 5, 2014 • 2:30 PM - 4:10 PM, Commonwealth A Coordinated Session, H7

Causal Modeling When Covariates are Measured With Error

Organizer and Session Chair: John Lockwood, Educational Testing Service, Pittsburgh, PA

Surrogate Balancing Scores: Matching on Covariates With Measurement Error

John R. Lockwood and Daniel F. McCaffrey, Educational Testing Service, Pittsburgh, PA

This paper considers strategies for using a surrogate, a variable measured with error, to construct an unbiased matching estimator for causal effects. It demonstrates that required conditions are unlikely to hold in applications. It argues that inverse probability weighting provides a more viable strategy for correcting for covariate measurement error.

Implications of Measurement Error on Covariate Selection for Causal Inference

Peter Steiner and Yongnam Kim, University of Wisconsin, Madison, WI

Estimating causal effects from observational studies via propensity score methods requires the reliable measurement of all confounding covariates. If they are measured with error bias reduction is attenuated, the inclusion of instrumental variables in the propensity score model amplifies remaining bias, balance checks lack power, and moderation effects are flawed.

Applying Multiple Imputation Using External Calibration to Propensity Score Estimation

Elizabeth Stuart and Yenny Webb-Vargas, Johns Hopkins, Baltimore, MD

Methods for handling covariate measurement error in propensity score estimation have not been widely investigated. We consider an approach that uses "Multiple Imputation using External Calibration" to correct for the measurement error. The approach works well in simulations, and we illustrate it using a study of early intervention for autism.

SIMEX for Weighting and Matching Applications With Error-Prone Covariates

Daniel McCaffrey and J.R. Lockwood, Educational Testing Service, Pittsburgh, PA

We propose Simulation-Extrapolation (SIMEX) for estimating causal effects or correcting for non-response using error-prone covariates. SIMEX adds more error to error-prone covariates, uses standard techniques on those data, and extrapolates to the value where there is no error. We show SIMEX mitigates bias but obtaining the correct extrapolation is challenging.

Saturday, April 5, 2014 • 2:30 PM - 4:10 PM, Commonwealth B Paper Session, H8

Inferences About Student and Teacher Preparation Derived From Measures of Student Learning

Session Chair: Matthew Gaertner, Pearson, Austin, TX

Measuring College Readiness in Middle School

Matthew Gaertner and Katie McClarty, Pearson, Austin, TX

This paper introduces a college-readiness index for middle school students. Using principal components analysis, we synthesized various middle-school indicators into six factors (achievement, behavior, motivation, social engagement, family characteristics, and school characteristics). Results suggest a diverse array of factors – most notably motivation and behavior – contribute substantially to college readiness.

Differential Prediction and Validity of AP for Student Subgroups

Minji K. Lee, UMass Amherst, Minneapolis, MN

The current study examines whether AP exam scores predict the first year GPA and second year retention differently for different groups of ethnicity, gender, parental education level, and language group, controlling for high-school-level variables using hierarchical linear modeling.

Investigating College and Career Readiness Through CCSS Aligned Tests

Thakur B. Karkee, Winnie K. Reid, Measurement Incorporation, Durham, NC; and Steven J. Aragon, Winston-Salem State University, Winston-Salem, NC

This study creates academic and demographic composites as indicators of College-Career Readiness (CCR) where state standards for large-scale assessments are aligned to CCSS. With success on high school exit exams as a surrogate for CCR, the results should illuminate the relationship between academic achievement, gender, race, SES and CCR.

Validating Measures of Teaching: Classroom Observations and Student Growth Percentiles

Andrea A. Lash, WestEd, San Francisco, CA; Benjamin Hayes, Washoe County School District, Reno, NV; Loan Tran, Min Huang, and Mary Peterson, WestEd, San Francisco, CA

Policymakers are turning to observations to evaluate teachers' instructional practices and to the Student Growth Percentile Model to evaluate teachers' effects on learning. Neither method has strong evidence of validity for high-stakes decisions. This validity investigation examined the internal structure of observation scores and their relationship to student learning.

Saturday, April 5, 2014 • 2:30 PM - 4:10 PM, Commonwealth C
Paper Session, H9

Bifactor and Multidimensional Models

Session Chair: Hyesuk Jang, Michigan State University, East Lansing, MI

Bi-Factor MIRT Equating for Testlet-Based Tests

Guemin Lee, Yonsei University, Seoul, Republic of Korea; Won-Chan Lee, Michael J. Kolen, University of Iowa, Iowa City, IA; In-Yong Park, Korea Institute for Curriculum and Evaluation, Seoul, Republic of Korea; Dong-In Kim, CTB/McGraw-Hill, Denver, CO; and Ji Seung Yang, University of Maryland, College Park, MD

The purposes of this study are to present bi-factor multidimensional item response theory (BF-MIRT) equating procedures for testlet-based tests and to investigate relative accuracy of them. Eight equating methods (both true- and observed-score) based on dichotomous IRT, polytomous IRT, testlet response model, and BF-MIRT are compared with target equipercentile equating.

Parameter Estimate Bias When Violating the Orthogonality Assumption: Bifactor Model

Chunmei Zheng, Pearson-Always Learning, Coralville, IA and Neal Kingston, University of Kansas, Lawrence, KS

A limitation of a bifactor model is the orthogonality assumption that proposes no correlations among the associated subdomains. To force correlated factors to be orthogonal can lead to distorted estimates. The purpose of this study, therefore, was to investigate the parameter estimate bias of different levels of orthogonality violation.

Bayesian Extended Two-Tier Full-Information Item Factor Analysis Model

Ken A. Fujimoto, University of Illinois at Chicago, Chicago, IL

An extended version of the two-tier full-information item factor model is presented. The model accommodates multiple primary (i.e., substantively meaningful) dimensions, correlated secondary (i.e., nuisance) dimensions, and higher-level person clustering. Based on the analysis of real rating data, the model displayed better predictive performance than other bifactor models.

Exploring the Examinee Locations Using Multidimensional Models Under Distributional Assumptions

Hyesuk Jang and Mark Reckase, Michigan State University, East Lansing, MI

This research investigates the estimation of examinee locations using the multidimensional latent trait models under different distributional assumptions. The effects of the distributional properties are theoretically explained and evaluated under the distributional assumptions using a simulation study. Distributional comparisons are provided and the implications related to practical issues are discussed.

Saturday, April 5, 2014 • 2:30 PM - 4:10 PM, Commonwealth D Paper Session, H10

Q Matrix Issues

Session Chair: Robert Henson, The University of North Carolina at Greensboro, High Point, NC

Minimal Requirements for Estimation of Q-Matrix and Item Parameters in CDMs

Sam Ye and Jeff Douglas, UIUC, Champaign, IL

CAT is an efficient format for estimation of ability profiles in cognitive diagnosis. In this research, we consider calibration of the Q-matrix values and item parameters of new items in the CD-CAT setting. Indices are proposed that generalize the global information index used in IRT-CAT.

A Case Study of Optimizing Q-Matrix in Cognitive Diagnostic Assessment

Cong Chen, Jinming Zhang, and Shenghai Dai, University of Illinois at Urbana-Champaign, Champaign, IL

This paper proposes a six-step procedure based on the fusion model and general diagnostic model to systematically optimize Q-matrix under real data settings. Results on a case study show such procedure helps to effectively reduce the subjectivity of Q-matrix construction and to improve Q-matrix performance for making cognitive diagnostic inferences.

The Role of Q-Matrix Design in Diagnostic Assessment

Matthew J. Madison and Laine P. Bradshaw, University of Georgia, Athens, GA

The Q-matrix precedes and supports any inference resulting from the application of a diagnostic classification model. This study investigates the effects of Q-matrix design on classification accuracy and reliability for the log-linear cognitive diagnosis model. Results indicate that Q-matrix design has a substantial effect on classification accuracy and reliability.

Dimensionality for Diagnostic Classification Models: Computing a "Scree Plot"

Robert A. Henson, The University of North Carolina at Greensboro, High Point, NC

The development of the q-matrix is among the most critical steps in DCMs. This paper describes an exploratory approach that first develops a scree plot to determine the number of attributes needed and then defines the corresponding Q-matrix.

Saturday, April 5, 2014 • 2:30 PM-4:10 PM, Millennium Hall
Graduate Student Poster Session, H11

Graduate Student Issues Committee
Lisa Beymer
Allison Chapman
Ian Hembry
Jason Herron
David King
Xiao Luo
Ting Wang

Poster Schedule

Saturday, April 5, 2014
9:40 a.m.-10:40 a.m.........................Setup Poster Presentation
2:30 p.m.-4:10 p.m. ....Author Present at their Poster
4:10 p.m.-7:00 p.m. .... Presenters Remove their Poster

Poster #1

Similarity-Measures for Nonparametric Examinee Classification

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

A recently introduced nonparametric approach to cognitive diagnosis classifies examinees into latent classes using distance measures between observed and ideal response patterns for given attribute profiles. This study introduces several similarity-measures as well as the conditions under which these measures provide more accurate examinee classifications for the nonparametric approach.

Poster #2

Poster #3

Educational Professionals' Understanding of the Concept of Validity

Katharine Bailey, Christine Merrell, and Peter Tymms, Durham University, Durham, United Kingdom

With shared responsibility for promoting valid interpretations of data, it is critical that assessment providers support teachers with meaningful guidance. Creating effective guidance must rely on gaining knowledge of what teachers understand validity and valid interpretations to mean. This study will explore UK teachers' understanding of validity issues.

Poster #4

Teacher Perceptions of the Washington State Teacher Evaluation System

Jessica L. Beaver and Brian French, Washington State University, Pullman, WA

Teachers' agreement and self-efficacy toward the skills measured on the new teacher evaluation system in the state of Washington were examined. General descriptive information along with factor analysis and latent mean comparisons on key demographic variables are presented. Understanding teachers' perceptions provides insight to the acceptance of these new systems.

Poster #6

Latent Interactions in MIRT for the Prediction of Reading Performance

Janine Buchholz, German Institute for International Educational Research, Frankfurt, Germany

A multidimensional item response theory model containing an interaction term was applied onto reading components test data in order to account for the theoretically assumed non-compensatory relationship between these components. Findings suggest that the model is more appropriate than the alternative standard model usually being employed.

Poster #7

Missing Not at Random: A Cause of DIF?

Kevin Cappaert and Yao Wen, University of Wisconsin at Milwaukee, Milwaukee, WI

Non-response is commonplace in practice. Tendency toward non-response can be thought of as a secondary dimension independent of the primary ability being estimated. A simulation will be conducted to test the effect of treating non-response as incorrect as a cause of DIF.

Poster #8

Exploring the Variability in Annual School Rankings

Allison Chapman and Don A. Klinger, Queen's University, Kingston, Canada

Important educational decisions are made based on schools' annual rankings on large-scale assessments (LSAs). Prior to such use, we must determine if the levels of academic performance on these LSAs within individual schools are stable over time. We found schools' LSA results differed across years, resulting in inconsistent annual rankings.

Poster #9

The Use of Cognitive Diagnostic Models With a Hierarchical Structure

Yi-Ling Cheng, Mark Reckase, Kelly Mix, Elizabeth Cook, Michigan State University, East Lansing, MI; and Susan Levine, University of Chicago, Chicago, IL

Three different CDM models were applied to data on map reading for young children. Fit of the models was compared and attribute profiles were estimated. We also test what would be a sufficient number of items to present a hierarchical structure in CDM models.

Poster #10

Effect of Ability Distributions in IRT Observed Score Equating

Jinah Choi, Won-Chan Lee, and Michael Kolen, University of Iowa, Iowa City, IA

This paper investigates the effect of distribution of ability on IRT observed score equating. Various combinations of ability distributions (e.g. a standard normal distribution and an uniform distribution) and characteristics of test forms, such as form difference or difficulty, are compared through simulation study with many replications.

Poster #11

Polytomous Extension of the DINA Model

Meng-ta Chung, Teacher's College, Columbia University, New York, NY

This study proposes a polytomous extension for the DINA model using the GPCM. The data is simulated using R, and the parameter estimation is carried out using WinBUGS.

Poster #12

Parameter Drift Methodology and Operational Testing Application

Amy K. Clark, University of Kansas, Lawrence, KS

The proposed research seeks to provide a comprehensive review of the item parameter drift literature and relevant methodology. A synthesis of findings across studies will be provided and implications for operational testing programs highlighted. Finally, an examination of item parameter drift for a statewide testing program will be examined.

Poster #13

Defining Q-Matrices of TIMSS 2007- 2011 Under Cognitive Diagnosis Modeling

Derya Evran and Jimmy de la Torre, Rutgers University, New Brunswick, NJ

This paper uses deterministic, inputs, noisy, "and" gate (DINA), and generalized deterministic, inputs, noisy, "and" gate (G-DINA) models in Trends in International Mathematics and Science Study (TIMSS) 2007 and 2011. Different Q-matrices are created with using the same attribute patterns across years to show the attribute similarity within TIMSS assessments.

Poster #14

Multidimensionality Assessment for the Variable Compensation Model

Yin Fu, University of South Carolina, West Columbia, SC

The variable compensation model subsumes both the compensatory and non-compensatory multidimensional IRT models. Holding the other parameters fixed, increasing the noncompensation decreases the reliability and difficulty dramatically, making such exams rare in practice,but not in simulation studies. The performance of DIMTEST is investigated after properly adjusting the item parameters.

Poster #15

Mending, Bending, Breaking: When Evidence Collides With Theory

Andrew Galpern, University of California at Berkeley, Berkeley, CA

This paper provides examples of the collisions between evidence and theory using a series of learning progressions co-developed with middle school science teachers in a large and diverse urban school system. The presentation outlines the research and design methodology we employed the methods we used to test those claims, the often unexpected results, and our responses to the evidence.

Poster #16

Evidence-Centered Design in Large-Scale Assessments of English Language Learners

M. Fernanda Gándara, University of Massachusetts at Amherst, Amherst, MA; Peter Swerdzewski, Arlen Benjamin-Gómez, and Kristen Huff, Regents Research Fund, New York, NY

Evidence-centered design appears as a promising approach to meet the challenges that current test development for English language learners faces. However, the lack of knowledge around how to use it may delay and even threaten its promises. This work pretends to fill such an applied knowledge gap.

Poster #17

Quantifying the Effects of Imprecise Item Parameters in CAT

Emre Gonulates, Michigan State University, Haslett, MI

The effects of using less precise item parameters instead of more precise ones on the ability estimates, standard error of ability estimates and test lengths of a computerized adaptive test (CAT) are investigated. The current paper quantifies the effects of using imprecise item parameters on these outcomes.

Poster #18

Truncated Sequential Probability Ratio Test Using Generalized Partial Credit Model

Samuel H. Haring, University of Texas, Cedar Park, TX

The truncated sequential probability ratio test (Spray & Reckase, 1994) is a termination procedure used to make classification decisions regarding examinee's performance. This simulation study evaluated the functionality of the procedure under the generalized partial credit model with varying content balancing, test length, and precision of classification conditions.

Poster #19

Multidimensional Item Response Theory in Large- Scale Assessments

Lauren Harrell and Li Cai, UCLA, Los Angeles, CA

In this study, large scale assessments such as NAEP and PISA are evaluated for potential local dependence of items. Multidimensional item response theory models are proposed to adjust for any residual correlations, and plausible value imputation procedures using MIRT models are explored.

Poster #20

The Development of an Item Pool Quality Index for Computer Adaptive Testing

Ying-Ju Hsu, University of Iowa, Iowa City, IA and Chingwei D. Shin, Pearson, Iowa City, IA

This study investigates and compares the properties of real and optimal item pools under different CAT settings. Two factors are manipulated in this study, including item exposure control methods and the complexity of content constraints. The purpose of this study is to provide a CAT pool quality index for administration.

Poster #21

A Comparison of Exposure Control Procedures in CATs

Xueying Hu and Sung-Hyuck Lee, ACT, Iowa City, IA

Proposes a new item exposure control procedure in computer adaptive testing (CAT). Compares the method with Sympon-Hetter procedure and with no exposure control procedures. Discusses advantages of the new procedure in terms of measurement precision and test security criteria. Indicates the practical application of the procedure.

Poster #22

A Comparison of IRT-Based Methods for Detecting Speededness

Hojun Hwang, University of Washington, Seattle, WA

This study evaluates response-based speededness estimates against speededness estimates from response-time models using real data using data from 20,000+ respondents to a college-level math exam. Latent class membership and speededness points will be estimated under MCMC and marginal MLE methods, and computational times will also be compared.

Poster #23

Empirical Q-Matrix Specification for Subsequent Test Forms: Further Investigation

Charles Iaconangelo and Nathan Minchen, Rutgers, State University of New Jersey, New Brunswick, NJ

The varsigma-squared method has been established as a viable method of data-driven Q-matrix specification for subsequent test forms (Huo & de la Torre, 2013). To increase its generalizability, this study evaluates the performance of the procedure across more realistic testing conditions, and investigates the range of optimal epsilon values used.

Poster #24

DIF for Spanish-Speakers in State Science Tests

Maria O. Ilich, University of Washington, Redondo Beach, CA

This large-scale study focused on ELLs using two DIF methods to analyze two years of fifth grade state standardized science tests. Item-type effects were consistent with previous research on minority group test performance. Differential bundling results indicated that DIF against ELLs was due to socioeconomic status and opportunity to learn.

Poster #25

Parameter Estimation Error When Ignoring Testlet Effects

Suk Keun Im and William Skorupski, University of Kansas, Lawrence, KS

The purpose of this study was to evaluate the effects on local item dependence (LID) on IRT parameter estimation. Data were simulated using Testlet Response Theory to produce LID. Results show that as testlet effects were increased, negatively biased discrimination, positively biased difficulty, and large random error in ability increased.

Poster #26

Item Response Theory Scale Linking Procedures in Multiple Mixed-Format Tests

Hyeon-Ah Kang, University of Illinois at Urbana-Champaign, Champaign, IL and Ying Lu, Educational Testing Service, Princeton, NJ

The study examines performance of concurrent calibration and a combination of concurrent calibrations with linking in multiple mixed-format tests. Simulation factors are linking and test design, equating coefficients, and item type. The simulation results show that concurrent calibration appears to work properly when multiple test forms need to be linked.

Poster #27

Examining Peer Human Rater Rubric Drift in Automated Essay Scoring

Abdolvahab Khademi, University of Massachusetts at Amherst, Amherst, MA

In measuring inter-rater reliability of an AES, the measurement rubrics by human raters may drift towards those of the AES, generating an inflated reliability. The present study attempts to examine whether such a drift occurs in measurement rubric scores generated by human raters when paired with AES vis-à-vis human raters.

Poster #28

Investigation of Item Properties Using the LLTM for Polytomous Items

Jinho Kim, Mark Wilson, and Karen Draney, University of California, Berkeley, CA

Among the Rasch family models, LLTM is useful to examine item properties to help predict success on a particular item. This study addresses how the LLTM can be applied to the Carbon Cycle data composed of newly developed polytomous items by comparing with PCM and RSM through the MRCML framework.

Poster #29

Constructing Attitude Scales with Category Boundary Variability

David R. King, James S. Roberts, Georgia Institute of Technology, Atlanta, GA; and Zane Blanton, University of North Carolina, Chapel Hill, NC

A fully Bayesian solution is given for the Law of Categorical Judgment that estimates category boundary standard deviations along with scale values, category boundaries, and stimulus standard deviations. Estimation accuracy was assessed through a recovery simulation and practical utility was examined by constructing an attitude toward gun control scale.

Poster #30

Missing Data Treatment When Estimating Growth With Educational Accountability Data

Jason P. Kopp and Sara J. Finney, James Madison University, Harrisonburg, VA

Missing data is common in longitudinal educational assessment, yet rarely addressed appropriately. Pre-post testing data were collected and initially missing posttest data were recovered via makeup assessments. Pre-post growth estimates obtained using the complete (i.e., including makeup) data will be compared to those obtained utilizing various missing data handling techniques.

Poster #31

PPMC-Analogous Approach to Testing Model Fit in Multilevel CFA

Megan Kuhfeld, UCLA, West Hollywood, CA and Li Cai, UCLA, Los Angeles, CA

This study investigates a PPMC-analogous (PP-PPMC: Lee & Cai, 2011) approach to detect misfit at a single-level of a two-level confirmatory factor analysis model. Results will be examined in the context of determining classroom-level factor structure in students' evaluations of teaching effectiveness surveys.

Poster #32

Diagnosing Student Mental Models Through Class Specific Scoring

Michelle M. LaMar, UC Berkeley, Berkeley, CA

Student conceptions are classified using an IRT mixture model in which classes are defined by ideal responses for the mental conceptions of interest. Student responses are scored against these ideals by class. We evaluate classification accuracy and parameter recovery in a simulation study. External validity is assessed using application data.

Poster #33

The Interaction Between the Source of Gender DIF on Mathematics

Ming-Chih Lan and Min Li, University of Washington, Seattle, WA

This study investigates an interaction effect between sources of gender DIF. Either content domain or item type along with cognitive demand acting as combined sources is examined. The existence of interaction effect is expected. The combination of multiple sources provides better pictures to describe gender DIF compared to individual sources.

Poster #34

Nonparametric IRT Parameters for Exploring Large Data Sets

Quinn N. Lathrop, University of Notre Dame, South Bend, IN

Using nonparametric "parameters" based on simple qualities of the kernel-smoothed ICCs provide quick exploration the items and are surprisingly close the 3PL MML estimates. In simulations, correlations with true parameters are comparable with an over 99% decrease in CPU time in the largest conditions. Implications for large data sets are discussed.

Poster #35

Examining "Slips" as Measurement Error Misclassification in Item Response Theory

Han K. Lee, University of South Carolina, Columbia, SC

The commonly used logistic IRT models do not consider the mistakes of examinees who know the answer but mark the incorrect response (e.g. slip). This contaminates the estimation of item parameters and underestimates examinee abilities. The Simulation and Exploration (SIMEX) technique is considered for handling this problem with off-the-shelf software.

Poster #36

Hypothesis Testing for the Adaptive Measurement of Unidimensional Individual Change

Jieun Lee and David J. Weiss, University of Minnesota, Minneapolis, MN

Two recent hypothesis testing methods to determine the significance of individual change using the adaptive measurement of change (AMC) were evaluated. Previous findings were extended and clarified. In simulation studies, bank characteristics and test length were varied both in conventional and adaptive tests. Type I error and power were examined.

Poster #37

A Comparison of Three Different Item Functioning Detection Methods

Philseok Lee and Seokjoon Chun, University of South Florida, Tampa, FL

This simulation compared three parametric DIF detection indices under Item Response Theory: Lord's Chi-square, Raju's NCDIF, and the Likelihood Ratio Test. We manipulated the percentage of DIF items (10%, 20%, 30%) and the type of DIF (Uniform and Non-Uniform). Results supported the superior performance of Raju's NCDIF.

Poster #38

Evaluation of Wald Test for Detecting DIF Under Multidimensional-IRT Models

Soo Youn Lee and Youngsuk Suh, Rutgers University, New Brunswick, NJ

The Lord's Wald test is evaluated and compared with the improved version using different estimation algorithms for DIF detection in the multidimensional-IRT framework. The original Wald test is implemented in BMIRT under the Bayesian approach and the improved version is implemented in IRTPRO under MLE approach.

Poster #39

Comparing Three Approaches for Estimating Parameters of Cognitive Attributes

Isaac Li, Yi-Hsin Chen, and Chunhua Cao, University of South Florida, Tampa, FL

A simulation study was designed to compare three methods for estimating cognitive attributes parameters: the linear logistic test model, the multiple regression method, and the crossed random-effects linear logistic test model. The effects of different population distributions and Q-matrices on accuracy and consistency of estimation were investigated as well.

Poster #40

Examining Black-White DIF Patterns on Mathematics Assessments

Xiangdong Liu and Catherine Welch, University of Iowa, Iowa City, IA

This study examined the previous black-white DIF research results from a national mathematic assessment. the result agree with previous DIF research in showing that black students outperform than white in algebra items. the results disagree with previous DIF research in showing that black students outperform than white in geometry items.

Poster #41

Yearly Math and Reading Skills Acquisition Trajectory of Elementary Students

Tanya Longabach, University of Kansas, Lawrence, KS

This study uses a multivariate multiple sample latent growth model to examine growth trajectories of elementary students in reading and writing in the course of one school year, as well as the similarities and differences in the trajectories of math and reading skills acquisition.

Poster #42

Setting Standard in Multidimensional Tests Using Bookmark Method

Xin Luo and Liyang Mao, Michigan State University, East Lansing, MI

Previous research on standard setting often assumes a unidimensional model. This research expanded current bookmark method to multidimensional tests and compared three variations: unidimensional IRT (UIRT) method, multidimensional reference composite (RC) method, and multidimensional IRT (MIRT method). MIRT was expected to recover the true cut score best.

Poster #43

Measuring Inter-Rater Reliability in Performance Assessment Items Using DIF

Tamara B. Miller, University of Wisconsin at Milwaukee, Milwaukee, WI

An alternative measure of inter-rater agreement for performance assessment items will be investigated. The proposed measure will utilize DIF analyses and doesn't require a fully crossed design. Empirical and simulated data will be used to examine if the proposed measure provides a more achievable and precise measure of inter-rater differences.

Poster #44

Accuracy of Multistage Estimates Across Discrete Stages of Testing

Kristin Morrison, HeaWon Jun, and Susan Embretson, Georgia Institute of Technology, Atlanta, GA

Multistage tests (MSTs) are increasingly used in a variety of testing settings due to their attractive qualities. In many settings, however, discrete estimates must be combined due to time lapses between stages. A simulation study was conducted and found that individual was better than information weighting of the discrete estimates.

Poster #45

Comparing Different Equating Methods Based on Equity Properties

Nese Öztürk Gübes and Hülya Kelecioglu, Hacettepe University, Ankara, Turkey

The performance of two item response theory (IRT) and three equipercentile equating methods were examined based on first-order equity (FOE) and second-order equity (SOE) criteria. The results showed that while IRT true-score equating method performed best in terms of preserving FOE, IRT observed-score equating method performed best in preserving SOE property.

Poster #46

IRT Model Comparison for Mixed Format Tests Equating

Seohong Pak, University of Iowa, Coralville, IA and Michael J. Kolen, University of Iowa, Iowa City, IA

This study investigate the main effect of twelve possible combinations of IRT models (1PL+GRM, 2PL+GRM, 3PL+GRM, 1PL+GPCM, 2PL+GPCM, & 3PL+GPCM) and scale transformation methods (Stocking and Lord & Haebara) on estimating parameters and IRT true equating and IRT observed equating for the mixed format test.

Poster #47

Using Latent Variable Models to Estimate Impact of Educational Interventions

Rachel Perlin, Ray Galinski, Jay Verkuilen, and Howard T. Everson, CUNY Graduate Center, New York, NY

Latent variable analyses (LVA) are used infrequently to estimate the impact of educational interventions. We demonstrate how LVA (e.g., missing data analyses, and IRT models) are used to evaluate pre-to-post-test gains in math achievement in an intervention that infuses mathematics in a middle school science curriculum.

Poster #48

Bayesian Simulation of Varying Response Scales Under Different Taxometric Procedures

Sarah M. Scott and Sarah Depaoli, University of California, Merced, CA

We examined instances within SEM in which factor loadings resulting from different item types are (in)accurately estimated. Confirmatory factor analyses with both categorical and continuous items exhibited extremely high levels of bias in both the frequentist and Bayesian frameworks employing default diffuse priors, suggesting cautious interpretation of substantive analyses.

Poster #49

A Comparison of Two Augmented Score Methods

Shuying Sha, University of North Carolina at Greensboro, Greensboro, NC

This simulation study is comparing the performance of two empirical Bayes estimation methods of subscore: Wainer et al's (2001) multivariate Bayesian estimation method and Harberman's (2005) "weighted average" or combined method. It is anticipated that when correlations between subscales are low, Harberman's method will perform better.

Poster #50

Multivariate Generalizability Analysis of Alternate Verbal Test for ELLs

AhYoung Shin and Hyung Jin Kim, University of Iowa, Iowa City, IA

This study evaluates the measurement precision of an alternate verbal format for young children with diverse linguistic backgrounds. The study provides validity evidence for using picture-based items as alternative to language-loaded items for measuring verbal reasoning abilities of young children including ELLs.

Poster #51

Multidimensional Learning Gains

Hyo Jeong Shin and Karen Draney, University of California at Berkeley, Berkeley, CA

This paper illustrates the use of multidimensional item response modeling to investigate the multidimensional learning gains. Using a large-scale science education assessment, we demonstrate 1) how to compare the learning gains across domains, and 2) how to classify the students onto the levels based on the learning progression framework.

Poster #52

Can an Assessment for Geometry Teachers Predict Student Outcomes?

Alan Socha, James Madison University, Gaithersburg, MD and Brett P. Foley, Alpine Testing Solutions, Orem, UT

There is a need for evidence that student geometry outcomes are related to measures of teacher quality. This study uses a three-level IRT model to evaluate relationships between teachers' knowledge of geometry, pedagogy, and student performance. Results will support stakeholders in identifying relations between teaching geometry and improving student achievement.

Poster #53

Comparing Different Model-Based Standard Setting Procedures

Ihui Su, Yong He, and Steven J. Osterlind, University of Missouri, Columbia, MO

Standard setting is a procedure to define the levels of competence and set cut scores. Model-based approaches were introduced to reduce subjectivity, but they have not been thoroughly explored the potentials in application. In this study, three model-based approaches (cluster analysis, latent class analysis, and mixture Rasch model) are compared.

Poster #54

The Simulating Study for Forced-Choice Data in DIF Detecting

Guo Wei Sun and Ching Lin Shih, National Sun Yat-Sen University, Kaohsiung, Taiwan

The purpose of this study was to understand DIF detecting performance under different type of force-choice data that can provide adequate statistical power for accurate DIF detection. The base DIF procedures were evaluated through applications to simulated data from thurstonian IRT model.

Poster #55

Evaluating Innovative Educational Interventions With State-Administered Proficiency Assessments: Validity Issues

Joshua M. Sussman, University of California at Berkeley, Berkeley, CA

This project explores the usefulness of state-administered proficiency assessments (SPAs) as outcome measures in summative evaluations of innovative educational interventions. Data simulation experiments grounded in an actual evaluation of the Learning Mathematics through Representations (LMR) intervention are used to develop an understanding of the validity of SPAs for evaluating innovative educational interventions.

Poster #56

DIF Analysis When Recoding the Polytomous Response as Dichotomous

Shuwen Tang, Wen Zeng, Cindy M. Walker, University of Wisconsin at Milwaukee, Milwaukee, WI; and Nina Potter, San Diego State University, San Diego, CA

The study focuses on the implications of recoding the polytomous response as dichotomous when using SIBTEST, as opposed to Poly-SIBTEST. We will conduct a simulation study to compare these approaches under different conditions, followed up with a feasibility study using real data obtained from a performance assessment of pedagogical efficacy.

Poster #57

Predicting Differential Item Functioning on Mathematics Items Using Multilevel SIBTEST

Juan A. Valdivia Vazquez, Brian F. French, Washington State University, Pullman, WA; and William H. Finch, Ball State University, Muncie, IN

Differential item functioning (DIF) is typically conducted in an ex post factor manner with no explicit hypotheses about which items differ between groups. In contrast, this study employed a new multilevel version of SIBTEST to account for complex sampling in a hypothesis driven framework to investigate DIF in mathematics items.

Poster #58

Trends in Learning Among Disadvantaged Student from 1999-2007

Marcus Waldman, Teacher's College, Columbia University, New York, NY

A novel extension to the DINA cognitive diagnostic model that incorporates student socioeconomic status is applied to the eighth grade TIMSS mathematics assessment. The goal is to discern to patterns in learning among the United States' most disadvantaged student under the No Child Left Behind era.

Poster #59

Combination of Ability Estimation and Classification in CAT

Keyin Wang and Liyang Mao, Michigan State University, East Lansing, MI

This study aims to find an item selection strategy to achieve the best classification accuracy and ability estimation recovery in a sequential probability ratio test (SPRT; e.g. Wald, 1947). Simulation was conducted for CAT with two-category-classification with the comparison of Fisher information, KL information and combined Fisher and KL.

Poster #60

Performance Comparisons Between Parametric and Nonparametric DIF Methods

Xiaolin Wang, Dubravka Svetina, Indiana University, Bloomington, IN; and Ou Zhang, Pearson, San Antonio, TX

Multiple methods have been developed to detect differential item functioning (DIF). We compare the performance of a parametric method (IRT-LR) and two nonparametric methods (Generalized Mantel-Haenszel and Poly-SIBTEST) in evaluating DIF. We investigate the impact of model-data misfit, sample size, group mean ability difference, and distributional shape of the populations.

Poster #61

Multidimensional Rasch Analysis of Indonesian Students' Performance on PISA 2012

Diah Wihardini and Mark Wilson, UC Berkeley, Berkeley, CA

This study examines the internal structure of the PISA 2012 mathematics assessment using multidimensional Rasch models. The Indonesian students' performance and the effects of several background variables on each assessed content domain are described. The findings provide important insights for the development and implementation of the new-yet-controversial national curriculum.

Poster #62

Inclusion of Covariates in Higher-Order Attribute Structure

Immanuel Williams and Charles Iaconangelo, State University of New Jersey, New Brunswick, NJ

Various attribute structures have been developed within the cognitive diagnosis model (CDM) framework. These include the higher-order attribute structure, which permits the estimation of general ability, ?, along with attribute mastery. This study evaluates the effect on attribute classification accuracy of the inclusion of covariates in the modeling of ?.

Poster #63

Generalizability Study of a Writing Assessment

Raffaela Wolf and Suzanne Lane, University of Pittsburgh, Pittsburgh, PA

Impact of construct-relevant and construct-irrelevant sources of variance on student performance assessment score inferences was examined through Generalizability Theory. The extent to which writing task scores can be generalized across prompts, raters and mode of writing was of interest in the current study.

Poster #64

Assessing Item Parameter Recovery for a Bifactor Model

Raffaela Wolf, University of Pittsburgh, Pittsburgh, PA

Accuracy of full-information item parameter estimation methods (Marginal Maximum Likelihood (MML) and Markov Chain Monte Carlo (MCMC) was examined for a Bifactor model in terms of Bias, RMSE, and SE under various test lengths and sample sizes.

Poster #65

Comparing MLR to Satorra-Bentler Estimation in MIMIC with Nonnormal Data

Pei-Chen Wu, Chunhua Cao, and Kim Eun Sook, University of South Florida, Tampa, FL

Distribution characteristics are critical in choosing statistical estimation methods. In a MIMIC model with one latent factor, six binary indicators, and one continuous covariate, both robust maximum likelihood (MLR) and Satorra-Bentler (SB) scaled test statistic were employed to conduct the likelihood ratio (LR) test and they performed equally well.

Poster #66

A Proposed Credibility Index (CI) in Peer Assessment

Yao Xiong, Hoi Suen, Wik-Hung Pun, Deborah Goins, and Xiaojiao Zang, Penn State University, State College, PA

Peer assessment of constructed-response assignments is promising in a massive-open-online-course (MOOC) environment, given the infeasibility of instructor grading the assignments of numerous students. However, there are concerns regarding the reliability/validity of peer ratings. A Credibility Index (CI) is proposed in this study to assess peer rater credibility.

Poster #67

An Empirical Investigation of Growth Models and School Accountability

Fan Yang and Stephen Dunbar, University of Iowa, Coralville, IA

This study aims to provide a conceptual overview as well as an empirical examination of various procedures for modeling student progress and evaluating school performance. Conditions of missingness and reliability of longitudinal data were suggested, and an empirical study was conducted to compare results of selected school accountability models.

Poster #68

Bayesian Estimation with Parallel Computing for 2PL Nested Logit Models

Ping Yang, Yong He, and Ze Wang, University of Missouri, Columbia, MO

We present a Bayesian hierarchical formulation for the 2PL-NLM and show that person and item parameters can be efficiently estimated. We provide the mathematical proof for applying customized sampling algorithms and parallel computing given conditional independence of model parameters. Studies based on simulation and empirical data will be conducted.

Poster #69

Automation of Any Graphical User Interface Software for Simulation Studies

Nedim Yel, Arizona State University, Tempe, AZ

Automation of statistical software for simulation studies can be challenging and time consuming. This paper introduces ways to automate potentially any statistical software which has graphical user interface by automatically sending mouse clicks and keyboard shortcuts to the program of interest.

Poster #70

Facilitating Standard Setting with Diagnostic Classification Modeling

Jiahui Zhang, Michigan State University, East Lansing, MI

A simulation-based method for standard setting is proposed by combining receiver-operating characteristic analyses and diagnostic classification modeling. Simulation studies showed reliability and efficiency compared with full-model analyses. This method produces flexible cut-scores with required specificity or sensitivity, increases the objectivity of standard setting, and reduces costs by lowering expert burden.

Poster #71

Comparing Dimensionality Assessment Approaches for Mixed-Format Achievement Tests

Mengyao Zhang, University of Iowa, Iowa City, IA

Determining test dimensional structures serves several important purposes. The widespread use of mixed-format tests in educational assessment poses challenges to analyzing test dimensionality. This study conducted a comprehensive review and comparison of dimensionality assessment approaches based on real data from several high school mixed-format achievement tests.

Poster #72

Exploring Children's Early Reading Achievement Gaps Using Hierarchical Linear Model

Mingcai Zhang and Lihong Yang, Michigan State University, Okemos, MI

This study investigates White and Hispanic children's early reading achievement gaps using hierarchical linear model. Our analysis shows that the gap exists between these two racial groups. Family socioeconomic backgrounds, race/ethnicity and gender are important factors that contribute to this gap.

Poster #73

Sample-Size Impacts on Standard Errors in Value-Added Models (VAMs)

Yan Zhou, Indiana University, Bloomington, IN and Edgar Sanchez, ACT, Iowa City, IA

This study focuses on the standard error estimates of VAM under different sample sizes. Current results indicated that sample sizes per teacher/school were negatively related to standard errors. A simulation analysis will be conducted to test the effects of sample sizes on standard errors in unbalanced and balanced designs.

Poster #74

Undergraduate Achievements are Predictive for Graduate-Level Performance in Computer Science

Judith Zimmermann, Kay H. Brodersen, Elias August, and Joachim M. Buhmann, ETH Zurich, Zurich, Switzerland

The indicative value of undergraduate achievements and which parts provide the best accuracy of graduate-level performance predictions was analyzed. Modern data mining techniques provided stable results: undergraduate achievements were highly relevant, particularly, the third-year grade point average. Admission rules were derived and the importance of the methods highlighted.

Saturday, April 5, 2014 • 4:20 PM - 6:00 PM, Washington A
Coordinated Session, I1

Equal Intervals and Vertical Scales: Figments of a Psychometric Imagination?

Organizer and Session Chair: Gary L. Williamson, MetaMetrics, Inc., Durham, NC

Testing the Quantitivity Hypothesis via the Trade-off Property

Alfred J. Stenner, MetaMetrics, Inc., Durham, NC

Several authors have expressed doubts that psychology can realize measurement as that term is used in physics. This paper asserts that measurement is possible, traces a roadmap to its realization and illustrates its attainment in the Lexile Framework for Reading.

The Interval Scale of Item Response Theory

David Thissen, The University of North Carolina at Chapel Hill, Chapel Hill, NC

This presentation discusses the bases of the claim that item response theory (IRT) yields scores on an interval scale, with special attention to the implications of the validity of that claim for vertical scaling, and the use of vertically scaled test scores in growth models.

Improving the Science Behind Vertical Scaling

Derek Briggs, University of Colorado, Boulder, CO

After examining some of the mixed messages about the properties of vertical score scales in the psychometric literature, three approaches are presented that could be used to evaluate the degree with which a vertical scale has equal-interval properties. Special emphasis is given to additive conjoint measurement as an implementable option.

Between the Ordinal and the Interval: Pliable Scales, Plausible Transformations

Andrew Ho, Harvard University, Cambridge, MA

I argue that either/or distinctions between ordinal and interval scales are unhelpful for most uses of educational test scores. Rather than confirming or disconfirming the interval nature of a score scale, I recommend and demonstrate the "pliability" of a scale and secondary statistics derived therefrom.

Saturday, April 5 • 4:20 PM–6:00 PM, Washington B
Invited Session, I2

An Introduction to Bayesian Modeling with Applications to Educational Measurement

Organizer: David R. King, Georgia Institute of Technology

Weighing Educational Evidence

Russell G. Almond, Florida State University, College of Education, Educational Psychology and Learning Systems

Evidence-centered assessment design (ECD) often represents our state of knowledge about a student's latent proficiency with a probability distribution. New evidence about the student's proficiency can be incorporated using Bayes' theorem. The weight of evidence, explored in this talk, is a measure of the change in that probability.

Bayesian Procedures for Evaluating Dimensionality Assumptions in Multidimensional Psychometric Models

Roy Levy, Arizona State University

This presentation describes the application of posterior predictive model checking, a Bayesian approach to assessing data-model fit, to the evaluation of dimensionality assumptions in psychometric models. Its advantages over frequentist approaches will be highlighted. Examples will be drawn from applications to multidimensional item response theory and Bayesian network models.

The Magic of Bayesian Applications to Educational Measurement

Lihua Yao, Defense Manpower Data Center at Monterey Bay

Bayesian methods and research on different models for educational and psychometric data are discussed. The BMIRT Toolkit suite of software is also discussed; this software utilizes Bayesian methods for Multidimensional Item Response Theory to improve score estimates for both paper-and-pencil and computer adaptive testing formats.

A Bayesian Model for Student Growth With Informative Dropout

Jeff Allen, ACT, Inc.

In the context of student growth models, informative dropout occurs when the probability of missing an assessment is related to a student's achievement level or rate of growth. To reduce bias associated with informative dropout, a Bayesian approach for jointly modeling the student growth and missing data processes is proposed.

Saturday, April 5, 2014 • 4:20 PM - 6:00 PM, Regency A
Paper Session, I3

Adaptive Item Selection

Session Chair: Xiaomin Li, The Hong Kong Institute of Education, Hong Kong, Hong Kong

Item Selection Methods in Variable-Length Cognitive Diagnostic Computerized Adaptive Testing

Xiaomin Li and Wen-Chung Wang, The Hong Kong Institute of Education, Hong Kong, Hong Kong

Within the context of variable-length cognitive diagnostic computerized adaptive testing (CD-CAT), we proposed a restrictive progressive item selection method and a restrictive proportional item selection method for item selection and exposure control. Their performance was evaluated with simulations and the results confirmed their effectiveness and feasibility.

Strand-Level Item Selection in CAT

Huijuan Meng, Pearson, Edina, MN; David Shin, and Yuehmei (May) Chien, Pearson, Iowa City, IA

This study compares the performance of "strand-level (SL)" and traditional item selection in CAT with respect to overall and domain theta estimation accuracy, bank usage, and constraint satisfaction. Three factors examined in the study were: correlation among true SL thetas, SL item selection starting point, and exam length.

Using Pretest Items to Fine-Tune Examinee's Interim Score in CAT

Changhui Zhang, Lisa Gawlick, Nancy Petersen, and Lingyun Gao, ACT, Inc., Iowa City, IA

In a computer adaptive test, pretest items are often embedded in the test but excluded from latent construct estimation. This study proposes the use of on-the-fly item parameter estimation and weighted interim scoring to obtain the information in examinees' responses to the pretest items.

Selecting Different Item Types in an Adaptive Test

G. Gage Kingsbury, Independent Consultant to NWEA, Portland, OR

Adaptive item selection models for use with homogeneous item pools are well established. However, when items differ in response structure, time demands, and scoring characteristics, an item selection procedure that accounts for these characteristics is needed. This study describes one such process, and compares it to existing adaptive testing models.

Saturday, April 5, 2014 • 4:20 PM - 6:00 PM, Regency B
Paper Session, I4

Measuring Teacher Effectiveness and Educator Performance

Session Chair: Daniel L. Murphy, Pearson, Austin, TX

Measuring Teacher Effectiveness With Cross-Classified Random Effects Models

Daniel L. Murphy, Pearson, Austin, TX and S. Natasha Beretvas, University of Texas, Austin, TX

This study examines the use of cross-classified random effects models (CCrem) and cross-classified multiple membership random effects models (CCMMrem) to estimate teacher effectiveness. Effect estimates are compared using CTT versus IRT and three models (i.e., multilevel model, CCrem, CCMMrem). Teacher effectiveness estimates are found to depend on scaling and modeling decisions.

Establishing Performance Standards to Support Educator Evaluation Systems

Elena Diaz-Billelo, Center for Assessment, Dover, NH, Erika Hall, Center for Assessment, Iowa City, IA, and Scott Marion, Center for Assessment, Dover, NH

Most educator evaluation systems necessitate the differentiation of performance in terms of 3 or more levels. The purpose of this paper is to advance technical understanding of the issues that influence the establishment of performance standards for educator accountability systems so that sound, defensible procedures can be defined.

Bias in Multilevel IRT Estimation of Teacher Effectiveness

Tongyun Li, Ming Li, Hong Jiao, and Robert Lissitz, University of Maryland at College Park, College Park, MD

The purpose of the present study is to investigate covariate inclusion in value-added models in relation to bias in teacher effect estimates. A simulation study is proposed to investigate this issue in the framework of value-added modeling using multilevel IRT models, specifically the three-level Rasch model with random teacher effects.

Comparisons Between Educator Performance Function-Based and Education Production Function-Based Teacher-Effect Estimations

Eun Hye Ham and Mark Reckase, Michigan State University, East Lansing, MI

This study aims to evaluate the feasibility of the educator performance function (EPERF)-based teacher effect estimation, by comparing with an existing education production function (EPROF)-based value-added model (VAM). The results from the two methods were compared and the model-fit of the EPERF was examined using a northern state's data.

Saturday, April 5, 2014 • 4:20 PM - 6:00 PM, Regency C1
Coordinated Session, I5

Spiraling Contextual Questionnaires in Educational Large-Scale Assessments

Organizer and Session Chair: Jonas Bertling, ETS, Princeton, NJ

Determining the Impact on Background Questionnaire Rotation on Population Modeling and Proficiency Estimation: A Simulation Study

David Magis and Christian Monseur, Université de Liège, Liege, Belgium

This paper presents a simulation study based on PISA 2003 data that investigates the impact of rotated background questionnaires (BQ) on the population modeling process, the drawing of plausible values and in fine, the related summary statistics. Based on these results questionnaire rotation should not be rejected as an alternative for collecting more information without increasing the testing time.

Imputing Proficiency Data Under Planned Missingness in Population Models

Matthias von Davier, Educational Testing Service, Princeton, NJ

This paper synthesizes the research on latent regression item response models for large-scale assessments, missing data, and imputations, as they apply to background questionnaires. It describes current best practices, as well as currently discussed solutions to cases that deviate from best practices. One of the main deviations is the use of data collections in which large amounts of observations are missing, completely at random, but by design.

Spiraling of Contextual Questionnaires in the NAEP TEL Pilot Assessment

Debby E. Almonte, Jonas P. Bertling, Janeen McCullough, Zhan Shu, and Andreas Oranje, Educational Testing Service, Princeton, NJ

This paper presents an application of questionnaire rotation to a recent NAEP large-scale field trial. An overview of the various methods that could be applied for spiraled questionnaires and how they compare with respect to various measures is given. Results for NAEP pilot data are compared and discussed.

A Review of Bayesian Imputation Methods Relevant to Background Questionnaire Rotation

David Kaplan, University of Wisconsin at Madison, Madison, WI

As with the rotation of cognitive assessments in surveys such as PISA and NAEP, the rotation of background questionnaires presents a complicated missing data problem. This presentation highlights recent developments in imputation theory from a Bayesian perspective and sets the groundwork for a systematic study of rotation of background questionnaires.

Saturday, April 5, 2014 • 4:20 PM - 6:00 PM, Regency C2
Paper Session, I6

Equating Methods and Consequences

Session Chair: Maria Bolsinova, Utrecht University/Cito Institute for Educational Measurement, Utrecht, Netherlands

The Impact of Equating on Detection of Treatment Effects

Youn-Jeng Choi, Allan S. Cohen, and Laura Lu, University of Georgia, Athens, GA

Equating makes it possible to compare performances on different forms of a test. Three different equating methods are examined for their impact on detection of treatment effects in an empirical data set.

Test Equating Using Prior Knowledge

Maria Bolsinova, Herbert Hoijtink, Utrecht University / Cito Institute for Educational Measurement, Utrecht, Netherlands; Anton Beguin, Cito Institute for Educational Measurement, Arnhem, Netherlands; and Jorine Vermeulen, Cito Institute for Educational Measurement / University of Twente, Arnhem, Netherlands

Test equating is used to make the results of examinations in different years comparable. To improve the quality of equating in non-equivalent group designs, we explore different ways of including prior knowledge based on experts' judgements and historical data in Bayesian estimation of the IRT models for test equating.

Investigating the Effects of Treating Missing Data on Vertical Scaling

Ahyoung Shin and Won-Chan Lee, University of Iowa, Iowa City, IA

The present study compares the utility of various imputation methods and other traditional ways of treating missing data with respect to the estimation of parameters in IRT. It also demonstrates how different ways of dealing missing data affect IRT vertical scaling and its interpretation about growth.

Factors Most Effecting Score Gaps in Equating

Stephen Cubbellotti, Fordham University, New York, NY and YoungKoung Kim, The College Board, New York, NY

This study investigates the impact of multiple factors on score gaps including: item difficulty, group ability, test length, equating methods, and the quality of anchor test. In addition, the impact of several combinations of the factors to evaluate the interaction among these factors using simulations within item response theory framework.

Saturday, April 5, 2014 • 4:20 PM - 6:00 PM, Commonwealth A Coordinated Session, I7

Exploring Some Solutions for Measurement Challenges in Automated Scoring

Organizer and Session Chair: Chaitanya Ramineni, ETS, Princeton, NJ

Rating Quality Definition Impact on Automated Scoring Engine Performance Depiction

Edward W. Wolfe, Pearson, Iowa City, IA, Peter W. Foltz, Pearson, Boulder, CO, and Leigh M. Harrell-Williams, Georgia State University, Atlanta, GA

This study compares agreement indices based on scores of automated scoring engines and humans from a pool of raters who scored all engine validation responses, allowing us to compare distributions of indices rather than point estimates. Results suggest advances in methodology for documenting the quality of automated engine scores.

Indices of Semantic Similarity for Automated Essay Scoring

Yoav Cohen and Anat Ben-Simon, NITE, Jerusalem, Israel

The current study examines the efficiency (validity) of six computer-generated semantic indices for an AES system for scoring texts in the Hebrew and Arabic languages. These include indices based on prompt-related vocabulary, semantic similarity, and Principal Component Analysis (PCA) of semantic similarities.

Reducing Automated Scoring Training Set Size With Alternative Sampling Methods

Shayne Miel, Measurement Incorporated, Durham, NC and David Vaughn, Measurement Incorporated, Carrboro, NC

The reliability of automated scoring of writing is largely dependent on the size and quality of the training data used to build the scoring models. This paper presents several methods for reducing the size of the training data while increasing the reliability of the predictions.

Using External Validity Criterion as Alternate Basis for Automated Scoring

Brent Bridgeman, Chaitanya Ramineni, Paul Deane, and Chen Li, Educational Testing Service, Princeton, NJ

This study investigates alternate weighting scheme for automated scoring models based on external validity criterion. Portfolio scores from the first year English composition course were used to build the models, and the feature weights were compared to weights from original models and to weights based on expert judgment.

Saturday, April 5, 2014 • 4:20 PM - 6:00 PM, Commonwealth B Coordinated Session, I8

Human Scoring Behavior and the Interplay Between Humans and Machines

Organizer and Session Chair: Jilliam Joe, ETS, Princeton, NJ

Features of Difficult-to-Score Essays

Tian Song, Pearson, Cincinnati, OH, Edward W. Wolfe, Pearson, Iowa City, IA, and Hong Jiao, University of Maryland, College Park, MD

This study predicts the difficulty of scoring an essay based on text features of the essay. The probability that human raters would assign scores that match true scores was modeled as a function of eight text features. Results indicate that text length and semantic typicality interact to predict scoring difficulty.

Human Rater Monitoring With Automated Scoring Engines

Mark Wilson, University of California, Berkeley, CA, Edward W. Wolfe, Pearson, Iowa City, IA, Peter W. Foltz, Pearson, Boulder, CO, Hyo Jeong Shin, University of California, Berkeley, CA and Tian Song, Pearson, Cincinnati, OH

This study compares rater effect classifications of human raters that are based on scores from human experts (HE) versus automated scoring engines (AE). Results suggest that AE and HE classifications are similar for rater bias, but that centrality rates are higher for AE and inaccuracy rates are higher for HE.

A Small-Scale Eye Tracking Study of Essay Rater Cognition

Jilliam Joe, ETS, Princeton, NJ, Isaac Bejar, ETS, Hamilton, NJ; Gary Feng, Mo Zhang, and Anita Sands, ETS, Princeton, NJ

This study used eye tracking technology to record the reading patterns of GRE(r) Analytical Writing Measure raters, and was used to guide their retrospective verbal reports. Preliminary results for experts suggest that they base their scoring decisions on the construct, but at times even they demonstrate less-than-ideal scoring processes.

Saturday, April 5, 2014 • 4:20 PM - 6:00 PM, Commonwealth C
Paper Session, I9

Cognitive Diagnostic Model Fit

Session Chair: Jonathan Templin, The University of Kansas, Lawrence, KS

Fit Indices' Performance in Choosing Cognitive Diagnostic Models and Qmatrices

Pui-Wa Lei, The Pennsylvania State University, University Park, PA and Hongli Li, Georgia State University, Atlanta, GA

This study examined the performance of fit indices in selecting correct cognitive diagnostic models (CDMs) and Qmatrices under different sample size conditions. The CDMs and fit indices from the R-package were investigated because they have not been evaluated extensively. Preliminary results suggest that the AIC index is most promising.

Modification Indices for Diagnostic Classification Models

Jonathan Templin, The University of Kansas, Lawrence, KS and Christy Brown, Clemson University, Clemson, SC

This study presents a method for detecting under-specification of a DCM and its Q-matrix, thereby indicating how the model could be modified to improve its fit to empirical data. This method is analogous to modification indices used in structural equation modeling, which are based on the score statistic.

Evaluation of Model Fit in Cognitive Diagnosis Modeling

Jinxiang Hu, David Miller, and Anne Corinne Higgins, University of Florida, Gainesville, FL

CDM model fit needs to be ascertained in order to make valid inferences. This study investigated the sensitivity of some model fit statistics including ?2, G statistics, -2LL, AIC, and BIC under different CDM settings inclusive of Q-matrix misspecification and CDM misspecification

Detecting and Characterizing Misspecification of Cognitive Diagnostic Models

Mark Hansen and Li Cai, UCLA, Los Angeles, CA

Here we examine the utility of Chen and Thissen's (1997) local dependence X^2 index for characterizing cognitive diagnostic model misfit. Through simulation study, the statistic is shown to be well-calibrated and sensitive to many types of misspecification. Use of the index is illustrated in an analysis of a large-scale mathematics assessment.

Saturday, April 5, 2014 • 4:20 PM - 6:00 PM, Commonwealth D
Paper Session, I10

Human Scoring Issues

Session Chair: Rianne Janssen, KU Leuven, Leuven, Belgium

Target Population Weights in Rater Comparability Scoring and Equating

Gautam Puhan, ETS, Princeton, NJ

When a constructed-response test is reused, rescoring at the current administration, of examinee responses from the previous administration provides the equating data to adjust for changes in the severity of the scoring. Theoretical and empirical examples (and results) show that the choice of target population weights for this equating is critical.

Features of Essay Responses Associated With the Accuracy of Human Scoring

Jie Chen, ACT, Inc., Iowa City, IA; Neal Kingston, University of Kansas, Lawrence, KS; Xuan Wang, Chris Richards, and Rick Meisner, ACT, Inc., Iowa City, IA

This study examines the scoring of writing items from a standardized achievement assessment. Score differences will be calculated for each scoring pair and item scoring accuracy will be compared to content coding and process characteristics of the rubric(s). Characteristics associated with more accurate scoring will be discussed.

Structuring Analytic and Holistic Judgments in Performance Assessments

Rianne Janssen, KU Leuven, Leuven, Belgium

A hierarchical classification model is used to structure analytic judgments on a performance assessment and to link the obtained structure to a holistic pass/fail rating. The model is illustrated with ratings of a performance task on oral communication competencies taken by 917 12th grade students.

Using Field-Test Information to Improve Post-Equating Accuracy

Lixiong Gu and Ying Lu, Educational Testing Service, Princeton, NJ

Two methods are examined to explore the use of field-test item statistics to augment the performance of post-equating. One uses the field test data as priors for post-equating item calibration. Another uses weighted average of post-equating item parameters and those from previous administrations to produce a conversion table.

Sunday, April 6, 2014 • 5:45AM–7:00 AM
Meet in the lobby of the Loews 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 Philadelphia. Meet in the lobby of the NCME hotel at 5:45AM. Pre-registration is required. Pickup your bib number and sign your liability waiver at the NCME Information Desk in the Loews Hotel, Regency Ballroom Foyer, anytime prior to race day.

The event is made possible through the sponsorship of:

- ACT
- Alpine Testing Solutions, Inc.
- American Institute of CPA's®
- American Institutes for Research®
- Applied Measurement Professionals, Inc.
- Buros Center for Testing
- Caveon™
- The College Board
- CTB/McGraw-Hill
- Data Recognition Corporation
- ETS®
- Graduate Management Admission Council
- Houghton Mifflin Harcourt
- HumRRO™
- LSAC®
- measured progress™
- Measurement Incorporated
- NBME®
- National Center for the Improvement of Educational Assessment, Inc.
- Pacific Metrics Corporation
- NCS Pearson, Inc.

Sunday, April 6, 2014 • 8:00 AM - 9:40 AM, Washington
Coordinated Session, J1

Scoring Issues for Next-Generation Performance Assessments: An Example from CBAL

Organizer and Session Chair: Randy Bennett, ETS, Princeton, NJ

A Ranking Method for Evaluating Constructed Responses

Yigal Attali, ETS, Princeton, NJ

This paper presents a comparative judgment approach for constructed response tasks based on ranking the quality of a set of responses. A prior automated evaluation of responses guides both set formation and scaling of rankings. A set of experiments shows that scores based on a single ranking outperform traditional ratings.

Generalizability of Results From Parallel Forms of Next-Generation Writing Assessments

Peter van Rijn, ETS Global, Amsterdam, Netherlands and Hanwook (Henry) Yoo, Educational Testing Service, Princeton, NJ

We report on results from a study of CBAL performance assessments in writing. The main question is the generalizability of results from parallel forms. Person-by-task interactions that are specific to a certain scenario context (and test form) might seriously hamper the meaning of test results.

Empirical Recovery of a CBAL Learning Progression for Linear Functions

Edith A. Graf, ETS, Princeton, NJ and Peter W. van Rijn, ETS Global, Amsterdam, Netherlands

Learning progressions are provisional models that require empirical validation. As part of earlier work, we used Wright maps to explore the empirical recovery of a linear functions learning progression. We extend that work with additional data, and compare the Wright map approach to a Bayes net approach.

Towards Meaningful Scores: Connecting Item Responses to Learning Progressions

Meirav Arieli-Attali, Educational Testing Service, Princeton, NJ

This paper describes our current work in developing categorical rubrics for scenario-based tasks and the characteristics of scores obtained via this method. Categorical rubrics classify student responses according to specific attributes linked to a hypothesized learning progression. These classifications have a potential classroom use in formative assessment settings.

Sunday, April 6, 2014 • 8:00 AM - 9:40 AM, Regency A
Paper Session, J2

Multidimensional and Two-Tier Models in Adaptive and Multistage Tests

Session Chair: Lihua Yao, Defense Manpower Data Center, Seaside, CA

Multidimensional Mastery Testing With CAT

Steven Nydick, Pearson VUE, Minneapolis, MN

The current study extends computerized mastery testing item selection algorithms and stopping rules to tests comprised of multiple dimensions. Rather than a single point separating masters from non-masters, multidimensional test require a boundary function. Conditions varied include type of boundary function, latent traits correlations, and item parameter structure.

Applying a Modified Multidimensional Priority Index for Variable-Length Multidimensional CAT

Ya-Hui Su, National Chung Cheng University, Chia-yi County, Taiwan

To achieve the same level of precision for examinees, stopping rules of precision were implemented with multidimensional priority index (Yao, 2013). However, Yao's method was proposed under between-item multidimensional framework. Therefore, a modified method was proposed for variable-length CAT under between-item and within-item multidimensional framework, and investigated through simulations.

Multidimensional Computer Adaptive Testing With Mixed Item Types

Lihua Yao, Defense Manpower Data Center, Seaside, CA and Haskell Sie, Pennsylvania State University, State College, PA

Performance of five multidimensional CAT item selection methods were compared using item pool of a mixture between multiple-choice and constructed-response items for mixed-format assessments. Ability recovery, each method's item preferences (simple- versus complex-structured items and location of maximum information), item exposure and item overlap rates were investigated.

CAT Using a Bifactor and Two-Tier IRT Models

Moonsoo Lee, CTB/McGraw-Hill, Monterey, CA

Bifactor and the two-tier IRT models in CAT were compared using simulation data. Two-tier model CAT worked well when two or more general factors were estimated properly. There was no significant effect of two-tier CAT algorithm for the specific factors, but it showed a significant effect for the general factors.

Sunday, April 6, 2014 • 8:00 AM - 9:40 AM, Regency B
Invited Session, J3

Diversity and Testing Issues

Organizer and Session Chair: Claudia P. Flowers, UNC Charlotte, Charlotte, NC

Growth models provide evidence of whether students are on-track to target proficiency levels, but there is a need to examine the impact for specific student subgroups. The purpose of this symposium is to present current research that explores the variations among growth models when applied to students with disabilities and English Learners.

Presenters:

Gerald Tindal, Director of Behavioral Research and Teaching, University of Oregon

In this presentation, different growth models will be presented for two different populations: (a) a transition matrix will be used to show growth (across years) for students with significant cognitive disabilities taking a statewide test, and (b) hierarchical linear models will be used to show (within year) growth for students with disabilities on curriculum-based measures. Both studies will highlight the need for complete and accurate data in making appropriate inferences on growth.

Joni Lakin, Assistant Professor, Auburn University, Department of Educational Foundations

This presentation uses state-level data to compare the behavior of four widely used growth models when applied to English Learner and non-EL students. Differences in the behavior of the models indicate that the choice of growth model can substantially impact inferences made about the academic progression of EL students.

Discussants:

Moderator: Martha Thurlow, Director of the National Center on Educational Outcomes, University of Minnesota

Derek Briggs, Professor and Program Chair, Research and Evaluation Methods, University of Colorado at Boulder

Michael J. Kolen, Professor, University of Iowa

Sunday, April 6, 2014 • 8:00 AM - 9:40 AM, Regency C1
Paper Session, J4

Through-Course and Interim Assessment: Growth and Comparisons With Summative Results

Session Chair: Thakur B. Karkee, Measurement Incorporated, Durham, NC

Comparing Through-Course and Across-Year Summative Assessment Growth Scores

Wwenhao Wang, University of Kansas, Lawrence, KS

The through-course summative assessment defined by the Race to the Top program as a system administered several times during the academic year for measurement of growth. This study compared the growth score calculated from this system with the growth score calculated from the more traditional cross-year summative assessments.

Comparing Student Growth Percentiles and Growth Norms: Interim/Summative Results

Joseph Betts, Houghton Mifflin Harcourt, Arlington Heights, IL

This research investigated Student Growth Percentiles and Growth Norms across two different types of data, summative and interim, for making judgments about student growth and comparing teacher rankings. Growth models were comparable, but results across the two types of data were highly inconsistent. A possible compensatory model is discussed.

Growth Prediction Using Interim Assessments

Lei Wan, Pearson, Coraville, IA, Ye Tong, Pearson, Iowa City, IA, and Jeff Barker, Gwinnett County Public Schools, Atlanta, GA

This study will explore the predictive validity of a district interim assessment, and to examine students' growth over the school year. The interim assessment data from a large school district will be used. Our ultimate goal is to provide insights about the use and design of interim assessments.

Sunday, April 6, 2014 • 8:00 AM - 9:40 AM, Regency C2
Paper Session, J5

Speededness

Session Chair: Weiling Deng, ETS, Princeton, NJ

Speededness Effect on Post-Equating Under CINEG Design

Min Wang, The University of Iowa, Iowa City, IA; Chunyan Liu, and Xiaohong Gao, ACT, Iowa City, IA

This study evaluated speededness effect on post-equating under CINEG Design. Speeded examinees were identified utilizing response time from a large-scale test on two content areas. Linear and IRT equating methods were applied to three chained forms for each content area. The practical impact of including/excluding speeded examinees was investigated.

Impact of Speededness on Item Performance and Population Invariance

Weiling Deng, Rui Gao, Neil Dorans, Chunyi Ruan, and Giunta Tony, ETS, Princeton, NJ

Using real data from a math test with timing information, this study evaluates whether speededness can be identified by examining IRT-based item performance. Results indicate that, for items affected by speededness, b-parameter varied for groups that reported different levels of speededness. Equating population invariance will also be examined.

Investigation on the Performance of Different Test Speededness Detection Methods

Haiqin Chen, University of Missouri, Columbia, MO; Feiming Li, and Hao Song, National Board of Osteopathic Medical Examiners, Chicago, IL

The performance of three test speededness detection methods including modified person fit index, lognormal response time (RT) model-based residuals, and speeded item response model with gradual change is compared via a simulation and empirical data, in an attempt to identify the relative strengths and weaknesses of these methods.

Modeling Speededness Using Survival Analysis

Carlos R. Melendez and Weiling Deng, Educational Testing Service, Princeton, NJ

The study uses survival analysis procedures to investigate the occurrence and timing of speededness, i.e. the point where the examinee's rate of item response time significantly increases. We expect that speededness varies as a function of age, gender, ethnicity as well as examinee ability and item difficulty.

Sunday, April 6, 2014 • 8:00 AM - 9:40 AM, Commonwealth A
Coordinated Session, J6

Longitudinal and Vertical Equating in the Australian National Assessment Program

Organizer and Session Chair: Goran Lazendic, The Australian Curriculum, Assessment and Reporting Authority, Sydney, New South Wales, Australia

Multistage Test Design Incorporating Vertical Scaling

Goran Lazendic, The Australian Curriculum, Assessment and Reporting Authority, Sydney, New South Wales, Australia and Raymond J. Adams, University of Melbourne, Melbourne, Australia

The feasibility of a multigrade and multistage test design in which testlets serve as vertical links between tests for different grades is investigated in the context of Australia's National Assessment Program. Such a design increases the robustness of the vertical scale and overall enhances these assessments.

Using Paired Comparisons to Equate Writing Performance Assessments

Steve Humphry and Joshua McGrane, The University of Western Australia, Perth, WA, Australia

The paper reports the results of a novel method of equating in which paired comparisons are used to longitudinally equate writing performance scales. The method involves the application of the Bradley–Terry–Luce model (Bradley & Terry, 1952; Luce, 1959) to analyse paired comparison data as part of a two-stage process of equating writing performance scales.

Removing Effects of Guessing in the Dichotomous Rasch Model

David Andrich, The University of Western Australia, Crawley, Western Australia, Australia

Guessed responses to multiple choice items produce bias in the Rasch model difficulty estimates. Recent research shows how such bias can be removed. Using vertical equating of a high profile national test, substantively important effects of removing guessing-induced bias on item difficulty and person proficiency estimates are demonstrated.

Equating and Scaling for Monitoring Student Achievement Over Time

Siek Toon Khoo and Yan Bibby, Australian Council for Educational Research, Camberwell, Australia

This presentation describes the equating procedures used in the Australian National Assessment Program – Literacy and Numeracy. The procedures used involve combining information obtained through vertical and longitudinal equating to place new assessment on the historic scale in order that achievement can be monitored over time across four grade levels.

Sunday, April 6, 2014 • 8:00 AM - 9:40 AM, Commonwealth B
Paper Session, J7

Mathematics Assessment

Session Chair: Nathan Dadey, University of Colorado at Boulder, Boulder, CO

Differentiated Assessment of Mathematical Competencies With Multidimensional Adaptive Testing

Anna Mikolajetz and Andreas Frey, University of Jena, Jena, Germany

The current study demonstrates the capability of multidimensional adaptive testing to enhance measurement precision for 11 subdimensions of mathematical literacy. A simulation study revealed substantially increased reliability coefficients for most of the content-related and process-related subdimensions of mathematical literacy.

Links Between Perceptions of Control, Persistence, Self-Esteem, and Mathematical Achievement

Stefanie R. McDonald and Rachel D. Upton, American Institutes for Research, Washington, DC

Based on data drawn from the Longitudinal Study of American Youth (LSAY), the differential performance of high-achieving and low-achieving African American students in mathematics is examined. The study employs latent growth modeling to investigate the influence of three intrapersonal traits: self-esteem, fate control, and persistence on the students' mathematics achievement.

Dimensionality at Multiple Levels: State-Level Diagnosis Using NAEP Mathematics

Nathan Dadey, University of Colorado at Boulder, Boulder, CO and Gregory Camilli, Rutgers, The State University of New Jersey, New Brunswick, NJ

Using multilevel multidimensional item response theory, we examine student- and state-level dimensions of academic achievement on the 2009 National Assessment of Educational Progress fourth grade mathematics assessment. The goal is to define and estimate state-level dimensions that can be examined in light of the influences of educational policy and practice.

Using Learning Progressions to Design Diagnostic Assessments in Mathematics

Leanne R. Ketterlin-Geller, Deni Basaraba, Pooja Shivraj, and Paul Yovanoff, Southern Methodist University, Dallas, TX

Learning progressions in mathematics provide the theoretical framework for development of five diagnostic tests for making instructional decisions. In this session we describe the analytic frameworks underlying the assessments including (a) the ordered multiple-choice structure that aligns with the learning progressions and (b) distractor analysis used to document students' misconceptions.

Sunday, April 6, 2014 • 8:00 AM - 9:40 AM, Commonwealth C
Paper Session, J8

Cheating Detection

Session Chair: Ardeshir Geranpayeh, University of Cambridge, Cambridge, United Kingdom

New Frontiers in the Application of IRT Based Cheating Indices

Ardeshir Geranpayeh, University of Cambridge, Cambridge, United Kingdom and Muhammad N. Khalid, Cambridge English Language Assessment, Cambridge, United Kingdom

This study will report on the findings of the application of two IRT based copying indices to investigate collusion in mixed format item type tests using both dichotomous and polytomous scored items. The study provides practical recommendations for test publishers in choosing the appropriate cheating indices for their investigation.

The K-Index With Exact Probability Model for Detecting Answer Copying

Hsiu-Yi Chao, Jyun-Hong Chen, and Shu-Ying Chen, National Chung Cheng University, Chiayi County, Taiwan

The K-index is a well-investigated method for answer copying detection. This study derived the exact probability model for the K-index to simplify its complex estimation process. Results showed that the K-index with the exact probability distribution can detect answer copying more effectively in both numerical and empirical study.

Power of Person-Fit Indices in Detecting Various Cheating Behaviors

Jiyoon Park, Yu Zhang, and Lorin Mueller, Federation of State Boards of Physical Therapy, Alexandria, VA

Examinees who take tests with the purpose of harvesting items may have item response patterns that are different from those of examinees who have item preknowledge. We evaluate the efficiency of selected person-fit indices in detecting aberrant response patterns in various cheating scenarios.

Sunday, April 6, 2014 • 8:00 AM - 9:40 AM, Commonwealth D Paper Session, J9

Item Calibration

Session Chair: Tammy Trierweiler, Prometric, Lawrenceville, NJ

Modeling Analysts' Judgments to Augment an Item Calibration Process

Carl H. Hauser, Northwest Evaluation Association, Beaverton, OR; Yeow Meng Thum, Wei He, and Lingling Ma, Northwest Evaluation Association, Portland, OR

Development and application of a model designed to augment the human review portion of evaluating field-test item performance is presented. The model, based on logistic regression, was shown to reduce the need for human reviews of all items by at least 70% with serious misclassifications of roughly 2%.

Evaluating Two Item Difficulty Equating Methods

Michael Walker and Usama Ali, ETS, Princeton, NJ

This paper concerns two methods used for equating item difficulty statistics. The first method involves linear equating and the other involves post-stratification. The paper evaluates these methods in terms of their standard error and robustness of equated item statistics using different number of test forms and sample size.

A Comparison of Scoring and Calibration Methods for Multipart Items

Yun Jin Rho, Pearson, Boston, MA and Hua Wei, Pearson, Saint Paul, MN

A multipart item refers to a set of items grouped together under a common stimulus. How can we deal with the multipart items in scoring and calibration? For this question, we evaluated different scoring and item parameter calibration methods to take into account the problems the multipart items can cause.

Scaling Item Parameters Using MCMC Estimation

Tammy J. Trierweiler, Prometric, Lawrenceville, NJ, Charles Lewis, Fordham University, New York, NY, and Robert L. Smith, American Institutes for Research, Washington, DC

MCMC estimation of item parameters provides uncertainty information that can be applied to test characteristic curves (TCCs). Matching MCMC estimates of TCCs for common items on two forms using the Stocking-Lord method provides uncertainty information about the resulting scaling. This approach is described and demonstrated using simulated data.

Sunday, April 6, 2014, 9:40 a.m.-10:00 a.m.
Regency Ballroom Foyer

Refreshment Break

Sunday, April 6, 2014 • 10:00 AM-11:40 AM, Washington
Invited Session, K1

10:00 AM - 10:50 AM

NCME Career Award Lecture

Session Chair: Joanna Gorin, ETS

Understanding Examinees' Responses to Items: Implications for Measurement

Susan Embretson
Georgia Institute of Technology

The cognitive processes and knowledge that examinees apply to solve items directly impact construct validity. Understanding these processes can improve item and test design, as well as provide a foundation for item generation and diagnostic measurement. Examples for several item types are presented, along with IRT models developed to estimate impact.

10:50 AM - 11:40 AM

Debate on Cognitive Approaches in Educational Measurement

Moderator: Joanna Gorin, ETS
Panel Members:
Russell Almond, Florida State University
André Rupp, ETS
Lorrie Shepard, University of Colorado at Boulder

Sunday, April 6, 2014 • 10:00 AM - 11:40 AM, Regency A
Invited Session, K2

Test Fairness: How Will the Revised AERA/APA/NCME Standards Affect Practice?

Organizer and Session Chair: Linda Cook, Educational Testing Service, Princeton, NJ

Fairness and Accessibility in the PARCC and Smarter Balanced Consortium Assessments: An Examination in Light of the Revised Standards for Educational and Psychological Measurement

Scott F. Marion, Center for Assessment, Dover, NH and Erika Hall, Center for Assessment, Iowa City, IA

Both PARCC and Smarter Balanced have committed to producing the fairest and most accessible assessments possible, while balancing political, practical, and technical constraints. We discuss the test development, accommodation and accessibility policies, and planned administration procedures of both consortia through the lens of the Fairness chapter of the revised Standards.

Fairness Standards and the ACT Assessment

Wayne J. Camara, ACT

This presentation will review the impact on the Fairness Standards on the development, use, and reporting of scores across large scale national assessments. ACT's processes for implementing equitable treatment, access to constructs, and validity of scores across test takers will be reviewed in relation to national testing programs.

Fairness Standards and College Board Assessments

Rosemary A. Reshetar, Gerald J. Melican, Sherri Miller, and Jay Happel, The College Board, Newtown, PA

The College Board assessments are regularly reviewed to ensure compliance with College Board and ETS Standards as well as the AERA/APA/NCME Standards for Educational and Psychological Testing. The new Standards afford an opportunity to look closely again. The presentation will address the steps and proposed future changes, if any.

Synthesis and Discussion

Joan Herman, UCLA/CRESST

The discussion will summarize common themes across the assessment examples presented, considering both the strengths and challenges in available evidence and design plans. The presentation also will identify promising approaches to meeting the revised fairness standards and conclude with consideration of opportunities and recommendations for research, policy and practice.

Sunday, April 6, 2014 • 10:00 AM - 11:40 AM, Regency B
Coordinated Session, K3

Detecting and Minimizing Sources of Construct-Irrelevant Variance in Performance Assessments

Organizer and Session Chair: Mark Raymond, NBME, Philadelphia, PA

Modeling Complex Performance Tasks
John Mattar, AICPA, Ewing, NJ and Matthew J. Burke, AICPA, Yardley, PA

Performance tasks (PTs) provide more realistic representations of the functions that professionals encounter in their field. These tasks are typically more complicated and expensive to develop. If organizations are interested in using them, the incorporation of task models should be considered to take the utmost advantage of PTs.

Practice Effects in a Performance Assessment of Physician Clinical Skills
Kimberly Swygert, National Board of Medical Examiners, Philadelphia, PA

We evaluated within-session and between-session score gains for 14,747 examinees who repeated, several weeks apart, a 6-hour performance test. Between-session score gains varied by education type and English language proficiency, and could be explained, in part, by within-session gains, indicating the presence of sizable practice effects.

Performance Tasks Comparisons to Multiple Choice and Essay Questions
Mark A. Albanese, National Conference of Bar Examiners, Madison, WI

Performance tests totaling 1874 examinees correlated modestly, but stronger with essay (r=0.41) than MCQ scores(r=0.32), supporting both construct validity and assessment of a distinct construct. Because performance tasks are expensive and time consuming, there are often fewer of them than other item-types, which present challenges in ensuring content validity.

Session Discussion: "Detecting and Minimizing Sources of Construct-Irrelevant Variance in Performance Assessments"
David Williamson, Educational Testing Service, Princeton, NJ

The discussant will identify elements common to each of the three presentations and summarize the lessons to be learned from each. The discussant will draw on the general validation frameworks proposed by Kane, as well as the validation strategies specific to performance assessment outlined by Messick.

Sunday, April 6, 2014 • 10:00 AM - 11:40 AM, Regency C1 Coordinated Session, K4

Linear Logistic Test Model and Item Variation

Session Chair: David Torres Irribarra, UC Berkeley, Berkeley, CA

Scale Score Properties for Performance Assessments Using Continuous Response Models

Benjamin Andrews, ACT, Iowa City, IA

In this paper, a continuous IRT model is used to estimate psychometric properties of scale scores such as CSEMs, reliability, classification consistency and classification accuracy for performance assessments with a large number of score categories. The effects of reducing the number of score categories are also investigated.

The Impact and Correction of Within-Template Systematic Variation

Quinn Lathrop, University of Notre Dame, Notre Dame, IN

Systematic variation within nested groups of generated items is a known but underexplored phenomenon. This proposal studies a model that explains within-template systematic variability in order to (1) improve ability and item parameter estimates and (2) bring insights about the educational process. Simulation studies are discussed.

Q-Matrix Misspecification in the LLTM Model With Random Item Effects

Yi-Hsin Chen, Isaac Li, Chunhua Cao, and George MacDonald, University of South Florida, Tampa, FL

A simulation study was designed to explore the effects of Q-matrix misspecification on parameter estimation in the LLTM model with random item effects. In addition, the impact of population distributions and Q-matrix density were also investigated. It is expected that the testing model performs best at the 2% under-misspecification.

The Ordered Linear Logistic Test Model (O-LLTM)

David Torres Irribarra, UC Berkeley, Berkeley, CA

This paper presents the Ordered Linear Logistic Test model (O-LLTM), a combination of the Linear Logistic Test model and the Ordered Latent Class model. The O-LLTM offers both straightforward respondent classification according to proficiency (for summative contexts) as well as explanatory analysis according to item features (for diagnostic purposes).

Sunday, April 6, 2014 • 10:00 AM - 11:40 AM, Regency C2
Paper Session, K5

Measurement Invariance and DIF

Session Chair: Jiyoung Yoon, Seoul Women's University, Seoul, Republic of Korea

DIF in CDM: Comparing Wald Test With Mantel-Haenszel and SIBTEST

Likun Hou, American Institute of CPAs, Ewing, NJ, Jimmy de la Torre, Rutgers, The State University of New Jersey, New Brunswick, NJ, and Ratna Nandakumar, University of Delaware, Newark, DE

Differential item functioning detection in cognitive diagnostic modeling has become of recent interest. This simulation study compares the performance of Wald test with Mantel-Haenszel and SIBTEST procedures in the context of the DINA model. Results show that the Wald test is comparable to or outperforms the other two procedures.

Differential Item Functioning in the G-DINA Model Using Standardized Differences Indices

Guaner Rojas, Universidad de Costa Rica, San José, Costa Rica, Jimmy de la Torre, Rutgers University, New Brunswick, NJ, and Julio Olea, Universidad Autonoma de Madrid, Madrid, Spain

Two test statistics for DIF detection based on the standardized differences between two IRFs for traditional IRT models are extended to the G-DINA model. A simulation study evaluated the Type I error and power of the statistics. Results showed that the indices performed well in detecting uniform and nonuniform DIF.

A Comparison Study of the Test Structures in TIMSS Tests

Jiyoung Yoon and Yoonsun Lee, Seoul Women's University, Seoul, Republic of Korea

The purpose of this study is to examine measurement invariance of the 2007 TIMSS tests among three countries: South Korea, U.S., and Singapore. The results showed that the multidimensional model with four factor model was supported for South Korea, U.S., and Singapore. Also, test fairness and validity analyses suggested that the TIMSS mathematics test was an appropriate instrument to compare students' achievement among the three countries.

Sunday, April 6, 2014 • 10:00 AM-11:40 AM, Commonwealth A Invited Session, K6

Dynamic Modeling Approaches

Session Chair: Markus Iseli, UCLA/CRESST

Network Psychometrics: Where Ising Meets Rasch
Günter Maris, Cito & University of Amsterdam

We show the Ising model (the cornerstone of statistical mechanics) to be a multidimensional Rasch model. The relationship between the Ising and Rasch model is shown to be mutually beneficial both for researchers in educational measurement, and for researchers in such fields as statistical mechanics, neural networks, etc.

Applications of Mixture Structural Equation Models with Regime Switching (MSEM-RS)
Sy-Miin Chow, Penn State University

Mixture structural equation model with regime switching (MSEM-RS) provides one possible way of representing over-time heterogeneities in dynamic processes by allowing a system to switch into and out of distinct change phases. Examples of MSEM-RS are illustrated using data from the Kindergarten Class of the Early Childhood Longitudinal Study.

Bayesian Analysis of Dynamic Item Response Models in Educational Testing
Xiaojing Wang, University of Connecticut

Based on time series testing data, a new class of item response models is proposed, which can be applied either to the full data or on-line, in case the real-time prediction is needed. The models are studied through simulations and applied to a reading test data collected from MetaMetrics, Inc.

On the Use of State-Space Models in Assessment and Instruction of Complex Tasks
Markus Iseli, UCLE/CRESST

Assessment of complex tasks coupled with coordinated instruction that fosters learning is challenging. Based on our previous research into games and simulations, we explore a proof-of-concept application that tracks learner performance over time and gives matching instructional feedback based on performance, with the overall goal of achieving set learning goals.

Sunday, April 6, 2014 • 10:00 AM - 11:40 AM, Commonwealth B Coordinated Session, K7

Leveraging Multiple Perspectives to Develop Technology-Enhanced, Scenario-Based Assessments

Organizer and Session Chair: Daisy Rutstein, SRI International, Santa Clara, CA

A Student Model for NAEP Science ICTs with CBAL Illustrations

Aaron Rogat and Lei Liu, ETS, Princeton, NJ

We outline a student model for science inquiry that is guiding the development of computer-based tasks for NAEP and we focus on a few selected elements of the student model that can be leveraged by technology. We illustrate some of these elements using examples from an ETS project called CBAL.

Developing Scenario-Based Science Inquiry Assessments

Daisy Rutstein, SRI International, Santa Clara, CA; Geneva Haertel, and Terry P. Vendlinksi, SRI International, Menlo Park, CA

This paper discusses the development of technology-enhanced scenario-based science assessments using an Evidence-Centered Design process. An ECD process is used to support the integration of the content that is being measured, the cognitive principles used when designing the items, and the use of technology.

Design of the SimScientists Assessments

Edys Quellmalz, Barbara Buckley, Mark Loveland, Matt Silberglitt, and Daniel Brenner, WestEd, Redwood City, CA

The SimScientists simulation-based assessment suites consist of curriculum-embedded formative assessments and end-of-unit summative benchmark assessments. The research-based designs integrate frameworks for model-based learning and evidence-centered design. Technology affordances allow representation of dynamic science systems "in action", and immediate feedback, customized coaching, and performance reports.

Best Practices in Developing Technology-Enhanced, Scenario-Based Science Assessments

George DeBoeur, Project 2061, AAAS, Washington, DC, James Pelligrino, University of Illinois at Chicago, Chicago, IL, and Robert Mislevy, ETS, Princeton, NJ

A panel of distinguished experts in large-scale, science assessment will address the degree to which science content and practice frameworks, technology enhancements, and cognitive principles have been integrated into the design of technology-enhanced, scenario-based science assessments. The role of ECD as an organizing framework will be discussed.

Sunday, April 6, 2014 • 10:00 AM - 11:40 AM, Commonwealth C Paper Session, K8

Items: Types, Parameters, Numbers

Session Chair: Jane Rogers, University of Connecticut, Storrs, CT

Performance of the PARSCALE G2 Statistic for Long Tests

Kyong Hee Chon, Western Kentucky University, Bowling Green, OH and Sandip Sinharay, CTB/McGraw-Hill, Monterey, CA

We examined the Type I error rates of the PARSCALE G2 statistic in a simulation study using sample sizes much larger than those considered in the literature. Our findings contradict the common belief that the statistic has reasonable performance for long tests.

An Effect of Item Type on Proficiency Distribution

Karen Draney, UC Berkeley, Oakland, CA and Hyo Jeong Shin, UC Berkeley, Berkeley, CA

This paper investigates whether the item type (i.e., structured interview, open-ended items, and forced-choice items) has an effect on proficiency distribution. Initial analyses using the Carbon Cycle project data indicate that as item types become easier to administer and to score, there is a reduction in the variance of the proficiency distribution.

The Impact of Homogeneity of Answer-Choices on Item Difficulty

Erkan Atalmis and Neal M. Kingston, CETE, Lawrence, KS

Haladyna and his colleagues (2002) proposed valid item-writing guidelines to construct high-quality MCIs. However, application of one guideline can lead to a violation of another guideline. This study investigates the impact of homogeneity of answer-choices on item difficulty when the plausible answer-choices with common solution errors of students are written.

Mitigating Effects of Item Parameter Estimation Error on Test Construction

Jane Rogers and Hariharan Swaminathan, University of Connecticut, Storrs, CT

The effect of IRT parameter estimation error on the psychometric properties of tests constructed using information functions and the utility of a correction to the information function that may mitigate the effects of estimation error were investigated. The correction factor was effective in reducing bias in the test information function.

Sunday, April 6, 2014 • 10:00 AM - 11:40 AM, Commonwealth D
Paper Session, K9

Providing and Evaluating 21st-Century Test Accommodations

Organizer and Session Chair: Stephen Sireci, University of Massachusetts at Amherst, Amherst, MA

Effects of Feedback and Revision for Students With Learning Disabilities

Elizabeth Stone, Cara Laitusis, Yigal Attali, and Carlos Cavalie, Educational Testing Service, Princeton, NJ

This presentation describes results from an experimentally designed field test in which 7th grade students with and without disabilities completed tests under both standard and feedback-and-revision conditions. We discuss results related to performance under the two conditions and affective reactions to correctness feedback.

Using Internal Structure Validity Evidence to Evaluate Test Accommodations

Stephen Sireci, Craig S. Wells, University of Massachusetts at Amherst, Amherst, MA; and Huiqin Hu, Data Recognition Corporation, Minneapolis, MN

Test accommodations promote access to educational assessments for many students. We analyzed data from statewide exams to compare tests taken under standard and accommodated conditions. Analyses of DIF, person fit, and dimensionality were conducted using replication and matching of proficiency distributions, which were important controls for properly interpreting the results.

Research and Development of Guidelines for Creating Accessible Computer-Based Assessments

Jennifer Higgins, Lisa Famularo, and Jessica Masters, Measured Progress, Newton, MA

This paper summarizes findings from a review of literature and state guidelines on audio and sign support, and cognitive lab research comparing different audio and sign representations of mathematics content. Findings were used to develop guidelines for how to appropriately represent mathematical notation and graphics in audio and sign form.

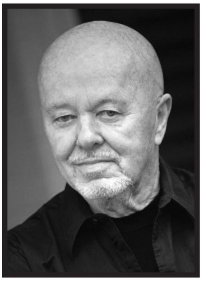
Evaluating Computer-Based Test Accommodations for English Learners

Katrina C. Roohr, Educational Testing Service, Princeton, NJ and Stephen G. Sireci, University of Massachusetts at Amherst, Amherst, MA

This study evaluated computer-based test accommodations for English Learners (ELs) on high school History and Math assessments. Using an experimental design, accommodation use and response time were evaluated for non-ELs and two EL groups using two direct linguistic accommodations. Results suggest new methods for future accommodation research.

Sunday, April 6, 2014 • 11:50 AM-12:40 PM, Washington
Invited Session, L1

Session Chair: Ronald Hambleton, University of Massachusetts, Amherst, MA



Criterion-Referenced Measurement: A Half-Century Wasted?

Jim Popham
UCLA

Fifty years ago, Robert Glaser introduced the world to criterion-referenced measurement. This new approach to educational testing had been triggered by the striking instructional successes he and others had seen while employing programmed instruction and teaching machines. Criterion-referenced testing, then, soon became regarded as a measurement strategy capable of triggering substantial improvements in students' learning. But has it? This presentation will supply a half-century answer to that question.

Sunday, April 6, 2014 • 11:50 AM-12:40 PM, Regency B
Invited Session L2

Session Chair: John Lockwood, ETS, Princeton, NJ

The Economic Value of Teacher Quality

Eric Hanushek
Stanford University

The substantial analysis of teacher effectiveness can now be linked to the economic returns both to students and to the overall economy. By linking estimates of the value-added of teachers to the increased achievement of students, it is possible to get direct estimates of economic outcomes. The analysis suggests very large returns to highly effective teachers.

Overall Electronic Board Session Chairs: Michael J. Culbertson, University of Illinois at Urbana-Champaign, Urbana, IL, and Marie Wiberg, Umeå University, Umeå, Sweden

Sunday, April 6, 2014 • 12:50 PM - 2:20 PM, Millennium Hall Electronic Board Paper Session 20

Equity and Equating

Session Chair: Jaime Peterson, University of Iowa, Iowa City, IA

Electronic Board Presentation #20.1A (Monitor 1)
A Meta-Analysis of Research on Repeated Attempts at an Assessment
Amanda A. Wolkowitz, Alpine Testing Solutions, Chesterfield

Numerous research studies discuss variables that influence an examinee's performance on multiple attempts at an assessment. Such variables include score differences between attempts, time between attempts, and the form used on each attempt. This meta-analysis synthesizes research on retake performance.

Electronic Board Presentation #20.2A (Monitor 2)
Summary of Vertical Scaling Systems Used in State Testing Programs
Jason Meyers, Pearson, Austin, TX and Ahmet Turhan, Pearson, Pflugerville, TX

Vertical scales provide useful student growth information to parents, teachers, and researchers. These systems are widely used in K-12 testing programs, and psychometricians must struggle with numerous decisions when constructing these scales. This paper summarizes the properties of the vertical scales currently in use through both descriptive and meta-analytic techniques.

Electronic Board Presentation #20.1B (Monitor 1)
Evaluating Equity at the Local Level Using Bootstrap Tests
YoungKoung Kim, Michael Chajewski, The College Board, New York, NY; and Lawrence T. DeCarlo, Columbia University, New York, NY

The present study proposes a method that evaluates equity using the bootstrap technique, which allows for a statistical test of equity at the local level without making any distributional assumptions. The proposed method is particularly helpful in high-stakes assessments where the determination of cut scores takes on great importance.

Electronic Board Presentation #20.2B (Monitor 2)
Multidimensional Item Response Theory (MIRT) Equating Procedures for Mixed-Format Tests
Jaime Peterson and Won-Chan Lee, University of Iowa, Iowa City, IA

This study aims to expand upon the Multidimensional Item Response Theory (MIRT) observed score equating literature in the context of mixed-format tests. Several real datasets from a large-scale testing program are used to illustrate a Full-MIRT equating and compare its results to traditional equipercentile, Unidimensional IRT, and Bifactor-MIRT equating procedures.

Sunday, April 6, 2014 • 12:50 PM - 2:20 PM, Millennium Hall Electronic Board Paper Session 21

Scaling and Growth

Session Chair: Jonathan Weeks, Educational Testing Service, Princeton, NJ

Electronic Board Presentation #21.1A (Monitor 3)
Developing an Interpretation Framework of Growth in Vertical Scaling
Meng Ye, Tao Xin, and Zhe Lin, Institute of Developmental Psychology, Beijing Normal University, Beijing, China

Considering the vagueness of the existing indexes of achievement growth, i.e. the properties of the vertical scale, in interpreting the individual growth, this paper employed the test characteristic function to express individual growth and constructed an interpretation framework of growth involving both the average grade growth and the individual growth.

Electronic Board Presentation #21.2A (Monitor 4)
Can the Currently Used Growth Models Measure "Growth"?
Shalini Kapoor and Catherine Jo Welch, University of Iowa, Iowa City, IA

Objectives of this research are to gauge the extent to which currently used growth models measure "growth", and propose "growth sensitivity" as a psychometric property of tests used for measuring student academic growth.

Electronic Board Presentation #21.1B (Monitor 3)
Evaluation of Single Group Growth Model Linking Method
Youhua Wei and Rick Morgan, ETS, Princeton, NJ

The single group growth model linking method uses common examinees between administrations to link test scores on different test forms. The study suggests that this new linking method is a potential alternative when there is a suspicion of the exposure of anchors in the common-item nonequivalent groups equating.

Electronic Board Presentation #21.2B (Monitor 4)
Implications of Construct Shift in Unidimensional Vertical Scaling
Jonathan Weeks, Educational Testing Service, Princeton, NJ

Interpretations of growth along a vertical scale are likely to be distorted when the data are modeled unidimensionally and the construct changes from one test to the next. This study examines growth interpretations and the identification of construct shift via an analytic and empirical comparison of separate and concurrent calibration.

Sunday, April 6, 2014 • 12:50 PM - 2:20 PM, Millennium Hall
Electronic Board Paper Session 22

Mapping and Alignment

Session Chair: Luz Bay, College Board, Dover, NJ

Electronic Board Presentation #22.1A (Monitor 5)
Mapping Importance Judgments in a Practice Analysis Using Correspondence Analysis
Andre F. De Champlain, Medical Council of Canada, Newtown, PA; Andrea Gotzmann, and Claire Touchie, Medical Council of Canada, Ottawa, Canada

Practice analyses, routinely undertaken to support of the development of professional licensure examinations, can be viewed as the precursor to the actual blueprint. This study provides a novel framework for assessing the impact of stakeholder group judgments in a national medical licensing examination practice analysis study using correspondence analysis.

Electronic Board Presentation #22.2A (Monitor 6)
Examining the Validity of Test-to-Curriculum Alignment Indices
Anne Traynor, Michigan State University, East Lansing, MI

Alignment index values are often presented as evidence that test content is representative of the performance domain defined by a written curriculum. Utilizing nine states' alignment, mathematics achievement test item, and teacher survey data, we examine the validity of a commonly-used alignment index.

Electronic Board Presentation #22.1B (Monitor 5)
Comparing Expert- and Student-Based Cognitive Models for an Algebra Exam
Zachary B. Warner, New York State Education Department, Albany, NY

The cognitive processes that students use to solve test items may be very different than those assumed by the developers. Think-aloud protocols were conducted to develop item-level cognitive models of task performance which were compared with the expert-based test specifications. This comparison yielded guidance for validation, development, and score reporting.

Electronic Board Presentation #22.2B (Monitor 6)
Investigating the Validity of the 2011 NAEP Writing Cut Scores
Luz Bay, College Board, Dover, NH, Susan Cooper Loomis, Consultant, Iowa City, IA, and Wonsuk Kim, Measured Progress, Dover, NH

The Body-of-Work standard-setting method was used to set achievement levels for the 2011 NAEP writing. The first validity investigation examines how the numerical cut scores relate to panelists' conceptual cut scores. The second analyzes the nature of the relationship between panelists' classifications of student work and the NAEP score scale.

Sunday, April 6, 2014 • 12:50 PM - 2:20 PM, Millennium Hall
Electronic Board Paper Session 23

Growth Percentiles/Validity Issues

Session Chair: Paul A. Westrick, ACT, Iowa City, IA

Electronic Board Presentation #23.1A (Monitor 7)
The Use of Student Growth Percentiles in School Accountability Assessment
Xin Lucy Liu, Shuqin Tao, Data Recognition Corporation, Maple Grove, MN; and Qiong Fu, Lehigh University, Bethlehem, PA

This study incorporates student background variables into the Student Growth Percentile (SGP) model for school accountability assessment. It reveals differential impact of student background factors on different quantiles of students. The effect of including these variables into the school accountability assessment via SGP is examined in real test data.

Electronic Board Presentation #23.2A (Monitor 8)
Comparing Student Growth Percentiles to Performance Level Change in Michigan
Dong Gi Seo, Michigan Department of Education, Lansing, MI and Adam Wyse, Michigan Department of Education, Minneapolis, MN

Several States have considered switching their models for measuring student growth. This study investigates how switching from a performance level change model to a student growth percentile model would impact assessment results in Michigan. Specific intention is given to similarities and differences at the student, school, and district levels.

Electronic Board Presentation #23.1B (Monitor 7)
Empirically-Derived Language-Minority Status and Validity Evidence
Do-Hong Kim, Richard Lambert, University of North Carolina at Charlotte, Charlotte, NC; and Diane Burts, Louisiana State University, Baton Rouge, LA

The primary purpose of this study is to identify unmeasured subgroups of language-minority children. The empirically derived subgroups are compared on social background variables and assessment results. This study also addresses the question of whether the use of latent classes can lead to more accurate conclusions of measurement invariance.

Electronic Board Presentation #23.2B (Monitor 8)
Validity Decay in STEM and Non-STEM Fields of Study
Paul A. Westrick, ACT, Iowa City, IA

This study used data from 62,212 students at 26 four-year institutions to examine the decay of validity coefficients for ACT scores and high school grade point average (HSGPA) over eight semesters in science, technology, engineering, and mathematics (STEM) fields and non-STEM fields of study.

Sunday, April 6, 2014 • 12:50 PM - 2:20 PM, Millennium Hall
Electronic Board Coordinated Session 24

Advances in Open-Source Assessment Technology: Benefits and Applications

Organizer and Session Chair: Marc Oswald, Open Assessment Technologies S.A., Esch-sur-Alzette, Luxembourg

Electronic Board Presentation #24.1A (Monitor 9)
A TAO Extension for Automated Generation of Job Skill Self-Assessment
Alexandre Baudet, CRP Henri Tudor, Luxembourg

This presentation highlights how TAO was used to automatically translate job descriptions into competency self-assessment for appraisal and for career guidance purposes in Luxembourg. Validity, reliability and satisfaction measures were used to show the efficiency of the innovative assessment algorithm and its associated computer-based tool.

Electronic Board Presentation #24.2A (Monitor 10)
Deploying TAO in Medical Education Using a Rapid Development Framework
Hollis Lai, University of Alberta, Edmonton, Alberta, Canada

In the realm of education, the adoption of open-platform technologies has been progressing at a staggering pace. However, assessment platforms are mostly proprietary technologies. TAO is the only open-platform for developing assessment administration software. In this presentation, we will demonstrate how TAO can serve as a viable alternative for meeting the testing needs in a large medical educational program focused on the use of formative assessment.

Electronic Board Presentation #24.1B (Monitor 9)
High-Quality Assessment Item Generation from Semantic Web Resources
Thibaud Latour, CRP Henri Tudor, Luxembourg

The creation of assessment items is an expensive task that does not scale well with manual approaches—this outcome highlights the need for the automatic generation of items. We present algorithms exploiting Linked Open Data and web resources, focussing on distractors quality using semantic similarity metrics and their empirical validation.

Electronic Board Presentation #24.2B (Monitor 10)
Open Source-Based Assessment Research and Delivery - Foresight or Folly?
Marc Oswald, Open Assessment Technologies S.A., Esch-sur-Alzette, Luxembourg, and Kate O'Connor, Breakthrough Technologies LLC, Chicago, IL

Thanks to open source initiatives such as TAO, a new paradigm has emerged that enables assessment specialists to save considerable time and money by leveraging advanced, freely accessible assessment software. This presentation will discuss the benefits of open source, and its impact on the field of educational assessment.

Sunday, April 6, 2014 • 12:50 PM - 2:20 PM, Millennium Hall
Electronic Board Paper Session 25

Cheating, Scoring, Scaling and Classifying in Multistage, CAT and Technology-Enhanced Assessment

Session Chair: William Lorie, Questar Assessment, Inc., Washington, DC

Electronic Board Presentation #25.1A (Monitor 11)
Application of a Scoring Framework for Technology Enhanced Items
William Lorie, Questar Assessment, Inc., Washington, DC

Technology enhanced items (TEIs) require students to interact with items in novel ways. Although templates exist for authoring TEIs, there is little guidance on how to score them. A TEI scoring framework was applied to response data for four forms, demonstrating the viability of partial credit scoring of TEIs.

Electronic Board Presentation #25.2A (Monitor 12)
Simultaneous IRT Scaling for Multistage Test
Deping Li, Yanlin Jiang, ETS, Princeton, NJ; and Chien-Ming Cheng, National Academy for Educational Research Preparatory, Taipei, Taiwan

This study is to examine how simultaneous scaling may be used in multistage test (MST) and how accurate the method is compared to those traditional online calibration methods such as Stocking-B, one EM cycle, and multiple EM cycles.

Electronic Board Presentation #25.1B (Monitor 11)
Using Full Bayesian Posterior Distribution to Classify Examinees in CAT
Liyang Mao and Xin Luo, Michigan State University, East Lansing, MI

The full Bayes procedure was compared with the confidence-interval, the sequential-probability-ratio-testing, and the Owen's sequential Bayes procedure, to determine which required fewer items for classification when the classification error rates were matched. The impact of different prior distribution on the full Bayes procedure was also examined.

Sunday, April 6, 2014 • 12:50 PM - 2:20 PM, Millennium Hall
Electronic Board Paper Session 26

Item Selection and Content Distribution in Adaptive Testing

Session Chair: Jing-Ru Xu, Michigan State University, East Lansing, MI

Electronic Board Presentation #26.1A (Monitor 13)
A Comparison of Fixed and Variable Content Distributions in CAT
Jing-Ru Xu, Michigan State University, East Lansing, MI, Xiao Luo, National Council of State Boards of Nursing, Chicago, IL, and Mark Reckase, Michigan State University, East Lansing, MI

We proposed two new designs of target content distribution that vary depending on different ability levels for a high-stakes licensure CAT. Different simulation studies were then applied to compare the results between one fixed and two variable target content distributions.

Electronic Board Presentation #26.2A (Monitor 14)
Comparison of Item Selection Methods in Adaptive Diagnostic Testing
Huiqin Hu, DRC, Plymouth, MN

Numerous studies have evaluated the item selection methods for computer adaptive tests. However, the few evaluate a statistical method when it is just one of the item selection criteria and is used to provide the diagnostic information. This study compares six statistical methods when non-statistical criteria are considered as well.

Electronic Board Presentation #26.1B (Monitor 13)
Item Bank Design for Constrained CAT With Non-Standard Normal Distributions
Xuechun Zhou, Tianshu Pan, and Ou Zhang, Pearson, San Antonio, TX

This study applied the p-optimal item pool design method for a constrained CAT program when non-standard normal distributions were anticipated. The results showed that ability distributions impacted items and pools characteristics evidently. Pool utilization also differed with respect to the number of items administered and item exposure rate.

Electronic Board Presentation #26.2B (Monitor 14)
Item Selection for Subsequent Sections in a CAT Test Battery
Lingyun Gao, Changhui Zhang, Nancy Petersen, and Lisa Gawlick, ACT, Iowa City, IA

This study evaluates three item selection methods for the later section of a two-section CAT test battery. Preliminary results reveal that selecting the initial item for Section 2 based on simulees' performance on Section 1 and then using EAP estimates as interim abilities provides the best estimation accuracy and efficiency.

Sunday, April 6, 2014 • 12:50 PM - 2:20 PM, Millennium Hall
Electronic Board Coordinated Session 27

Reporting Student Growth: Issues and Future Score Reporting Efforts

Organizer: Haifa Matos-Elefonte, College Board, New York, NY
Session Chair: Thanos Patelis, Center for Assessment, Dover, NH

Electronic Board Presentation #27.1A (Monitor 15)
Beyond a Single Estimate: Providing Insight and Information to Educators
John White and Nadja I. Young, SAS Institute, Cary, NC

Dr. John White, SAS Institute's Director of EVAAS, will illustrate how educators can benefit from statistically robust reporting that can be used in improving students' performance. In going beyond a single estimate of effectiveness, reporting based on value-added and growth metrics becomes much more powerful and useful to educators and policymakers.

Electronic Board Presentation #27.2A (Monitor 16)
Making Growth Score Reporting More Meaningful to Users
April L. Zenisky, University of Massachusetts at Amherst, South Hadley, MA and Ronald K. Hambleton, University of Massachusetts at Amherst, Amherst, MA

Recent efforts to report student growth using complex statistical models present unique reporting challenges, complicating ongoing efforts to effectively report performance. This paper focuses on three areas of particular interest in reporting: (1) strategies for reporting growth for individuals and groups, (2) highlighting score imprecision, and (3) reporting subtest scores.

Electronic Board Presentation #27.1B (Monitor 15)
On Interactive Media and Student Score Reporting
Adam VanIwaarden, University of Colorado, Boulder, CO, Damian Betebenner, Center for Assessment, Dover, NH, and Ruhan Circi, University of Colorado, Boulder, CO

Individual longitudinal score reports based upon interactive media permit "layers of meaning" to be placed in front of the viewer resulting in complex, yet simple to understand, reports capable of engaging stakeholders at various levels. Such reports can present a comprehensive view of student performance while simultaneously raising the level of stakeholder (e.g., parent, student, teacher) engagement around student data.

Electronic Board Presentation #27.2B (Monitor 16)
Reporting Growth Within a College Readiness System
Haifa Matos-Elefonte, College Board, New York, NY; Carol Barry, Amy Hendrickson, Kara Smith, College Board, Newtown, and Thanos Patelis, Center for Assessment, Dover, NH

The application of growth models to a suite of college readiness assessments has led to thoughts about how to display results of these models and issues that should be addressed when reporting to stakeholders. The second paper describes score reporting efforts from the perspective of researchers within a testing company.

Sunday, April 6, 2014 • 12:50 PM - 2:20 PM, Millennium Hall
Electronic Board Paper Session 28

Automated Scoring: Sampler of Applications

Session Chair: Qiwei He, ETS, Princeton, NJ

Electronic Board Presentation #28.1A (Monitor 17)
The Impact of Anonymization on Automated Essay Scoring
Mark D. Shermis, The University of Akron, Akron, OH, Sue Lottridge, Pacific Metrics Corporation, Lakewood, CO, and Elijah Mayfield, LightSide Labs, Pittsburgh, PA

This study investigated the impact of anonymizing essay text on the predictions made by AES. Essay data were analyzed under two conditions—in its original form and after processing by the Stanford NER which eliminates identifiable references. There were no differences on the score predictions made by two AES engines.

Electronic Board Presentation #28.2A (Monitor 18)
Combining Text Mining and IRT Estimates to Improve Test Efficiency
Qiwei He, Educational Testing Service, Princeton, NJ; Bernard Veldkamp, and Cees Glas, University of Twente, Enschede, Netherlands

This study presents a trial to combine text mining techniques and IRT estimates in a systematic framework. Two specific objectives were investigated: whether adding textual assessment could enhance the classification accuracy of examinees, and whether the standard error of estimates could be reduced compensating from a text prior.

Electronic Board Presentation #28.1B (Monitor 17)
Identifying Predictors of Machine/Human Reliability for Short-Response Items
Claudia Leacock, CTB McGraw-Hill, New York, NY and Xiao Zhang, CTB McGraw-Hill, Monterey, CA

We identify features of items that correlate with agreement in engine/human scoring. This is an extension of Leacock, et al (2013) by (a) increasing the pool of items from 41 to 570, (b) including both math and ELA, (c) greatly increasing the number of features studied.

Electronic Board Presentation #28.2B (Monitor 18)
Modeling Log Data from an Online Game Using Exploratory Approaches
Yuning Xu, Arizona State University, Tempe, AZ, Kristen E. DiCerbo, Pearson, Chandler, AZ, and Roy Levy, Arizona State University, Tempe, AZ

This study analyzed log data from an online role-playing game to predict outcomes using a longitudinal approach to data management and exploratory regression models. The results indicate strong predictive power early, with degradation over time. It is argued these methodological approaches may be gainfully employed in analyses of log files.

Sunday, April 6, 2014 • 12:50 PM - 2:20 PM, Millennium Hall
Electronic Board Coordinated Session 29

Analyzing Data from Collaborations: Structure, Models, Applications

Organizer and Session Chair: Alina von Davier, Educational Testing Service, Princeton, NJ

Electronic Board Presentation #29.1A (Monitor 19)
Models for Scoring Individual Performance in Team Collaboration
Peter, F. Halpin, New York University

Team work can be conceptualized as a sequence of actions taken by each member toward one or more overall goals. We analyse these sequences of actions using new IRT-like models that condition not only an individual's ability, but also the ability of their team members.

Electronic Board Presentation #29.2A (Monitor 20)
Investigating Collaborative Problem Solving with Extended Trialogue Tasks
Jiangang Hao, Alina A. von Davier, Lei Liu, and Patrick Kyllonen, Educational Testing Service

We present the initial results from Trialogue CPS (collaborative problem solving) project, which is designed to investigate the CPS in a game-like environment. In the game, two human participants collaborate via message boxes and interact with two virtual agents. Their communications and actions are analyzed to study the process of CPS.

Electronic Board Presentation #29.1B (Monitor 19)
Dynamic Bayesian Network Models for Peer Tutor Interactions
Yoav Bergner, Educational Testing Service, and Erin A. Waller, Arizona State University

A naïve hidden Markov model developed in an earlier study is refined for better accuracy of the inputs and extended for generalizability to complex interactions. We explore the capabilities of dynamic Bayesian networks (DBNs) for modeling the interplay between individual abilities, collaborative acts and performance outcomes.

Electronic Board Presentation #29.2B (Monitor 20)
Social Networks and Team Performance in Large Scale Online Role Playing Games
Mengxiao Zhu, Educational Testing Service

This study analyzes team performance and team assembly using a dataset from a popular U.S. based Massively Multiplayer Online Role-Playing Game (MMORPG), EverQuest II. Social network analyses and three measures of team performance are used to predict successful team outcomes.

Sunday, April 6, 2014 • 12:50 PM - 2:20 PM, Millennium Hall
Electronic Coordinated Paper Session 30

Revising Assessment Systems Around the World: Evolution, Revolution, or Both?

Organizer: Alvaro J. Arce-Ferrer, Pearson, San Antonio, TX

Session Chair: Avi Allalouf, NITE, Jerusalem, Israel

Electronic Board Presentation #30.1A (Monitor 21)
Adding an Essay-Writing Section to the PET and More
Avi Allalouf and Naomi Gafni, NITE, Jerusalem, Israel

PET is a test used for admission to higher education in Israel. Three changes were made in 2011: The PET has included a writing task; items with the lowest face validity and authenticity were excluded from the test; and three general scores (instead of one) were computed and reported

Electronic Board Presentation #30.2A (Monitor 22)
Linking Similar Tests with Different Content Standards: Anchor Purification Error
Alvaro J. Arce-Ferrer, Pearson, San Antonio, TX

This paper presents a formulation of expected anchor purification error in the context of test linking. Anchor purification error indexes expected total error due to handling item parameter drift (IPD) flags. The paper develops indexes to formally study anchor set purification decisions to link tests developed with different content standards.

Electronic Board Presentation #30.1B (Monitor 21)
On the Limits of Linking: Experiences from England
Anthony D. Dawson and Paul Newton, Institute of Education, University of London, UK

When operating at the limits of linking – e.g. during periods of curriculum and examination reform – the threat of encountering serious problems is high. To identify lessons for future reforms, this paper considers the two most serious examining crises of recent years in England, both of which involved examinations in transition.

Electronic Board Presentation #30.2B (Monitor 22)
The New SweSAT: Reasons, Revisions, and Results
Per-Erik Lyrén and Christina Wikström, Umeå University, Umeå, Sweden

This presentation describes the SweSAT, reasons for recent revisions, and results from initial validity studies. Some of the expected positive consequences of the revisions are yet to be confirmed, and the revisions also introduced some new challenges when it comes to score reporting and use, scaling and equating, and accommodations.

Sunday, April 6, 2014 • 2:30 PM - 4:10 PM, Washington Coordinated Session, M1

Developments in Statistical and Psychometric Modeling for International Large-Scale Assessments

Organizer and Session Chair: Leslie Rutkowski, Indiana University, Bloomington, IN

Modeling ILSA Data Through a Bayesian Paradigm

David Kaplan, University of Wisconsin at Madison, Madison, WI

Utilizing data from two cycles of PISA, this paper demonstrates Bayesian statistical inference via several examples. We specify several linear and multilevel models using data from PISA 2009, comparing the case where priors are not used to the case where informative priors based on results from PISA 2000 are used.

Alternative Models for Population Achievement Estimation

Matthias von Davier, ETS, Princeton, NJ

The proposed presentation proposes latent class analysis (LCA) as an alternative to the traditionally used principal component analysis for developing the conditioning model in large-scale assessments. LCA is used to identify one or more latent nominal variables that can be used to classify respondents with respect to their background characteristics.

Multilevel Modeling of Categorical Variables From International Large Scale Assessments

Carolyn Anderson, University of Illinois at Urbana-Champaign, Champaign, IL, Jee-Seon Kim, University of Wisconsin at Madison, Madison, WI, and Bryan Keller, Teachers College, Columbia University, New York, NY

Models and methods for analyzing discrete response variables are extended to the multilevel context, including some novel models that relax restrictive assumptions. Design weights to deal with unequal probability sampling is discussed and illustrated, and a method for handling missing data for multilevel data is presented.

Modeling Country-Specific Differential Item Functioning

Cees Glas, University of Twente, Enschede, Netherlands

The current paper examines improving the fit of measurement models in international assessments by using country specific item parameters to model DIF. Tests of model fit are used to establish that the two sets of items relate to the same latent variable, yet with different item parameters.

Sunday, April 6, 2014 • 2:30 PM - 4:10 PM, Regency A
Paper Session, M2

Calibration and Balancing Information

Session Chair: Yi Zheng, University of Illinois at Urbana-Champaign, Champaign, IL

The Ordered Informative Range Priority Index Method for Online Calibration

Yi Zheng and Hua-Hua Chang, University of Illinois at Urbana-Champaign, Champaign, IL

This paper proposes a new method ("OIRPI") for pretest item selection in online calibration in computerized adaptive testing. A simulation was conducted to compare OIRPI with existing methods under the 1/2/3PL models. Results show OIRPI effectively improves the calibration efficiency over the existing item selection methods in online calibration.

Viability of a Quasi-LOFT Approach for New Testing Programs

Huijuan Meng, Pearson, Edina, MN; Jennifer Davis, and Kathi Gialluca, Pearson, Bloomington, MN

This study investigated the viability of adopting a less-constrained Linear-on-the-Fly testing (LOFT) design for a newly developed testing program. LOFT and less-constrained LOFT (quasi-LOFT) exams were compared to a linear fixed-form exam with respect to ability estimation accuracy and classification accuracy under different research conditions.

A New Online Calibration Approach for Multidimensional Computerized Adaptive Testing

Ping Chen and Tao Xin, Beijing Normal University, Beijing, China

In view of the theoretical weakness of M-Method A (Chen & Xin, 2013), this study proposed a new MCAT online calibration method, named FFMLE-M-Method A, which incorporates the full functional MLE (Jones & Jin, 1994) into M-Method A with a goal to correct for the estimation error.

Balancing Information With Imbalanced Item Banks in Multidimensional CAT

Michael J. Culbertson, University of Illinois, Urbana, IL

In multidimensional Computerized Adaptive Testing (CAT), if item banks provide more information on one dimension than others, most item selection algorithms will over-emphasize the information-rich dimension. This study evaluates two proposed selection criteria based on mutual information that are intended to equalize ability estimate precision across dimensions.

Sunday, April 6, 2014 • 2:30 PM - 4:10 PM, Regency B
Paper Session, M3

Time Elements in Adaptive Tests

Session Chair: Shu-chuan Kao, NCSBN, Chicago, IL

Quantifying Item-Level Time Sensitivity for Items in Computerized Adaptive Tests

Shu-chuan Kao, Tony Zara, Pearson, Chicago, IL; and Ada Woo, NCSBN, Chicago, IL

The purpose of this study is to identify items exhibiting differential speededness. Indices are proposed to reflect the degree of time sensitivity based on item difficulties calibrated on groups with different item-level response-times. The proposed method can provide useful information for item review, standard setting, and item selection in CAT.

Is it Practical to Use Response Times in CAT?

Hong Qian, National Council of State Boards of Nursing, Evanston, IL, Mark Reckase, Michigan State University, East Lansing, MI, and Philip Dickison, National Council of State Boards of Nursing, Chicago, IL

An approximation estimation procedure was proposed to incorporate response times to estimate ability in CAT, as an alternative to MCMC in original research which takes too long for CAT to estimate ability after each item. The results showed this procedure reduced test length at the cost of increased estimation bias.

Using Response Times to Improve Estimation in CAT

Justin L. Kern, University of Illinois at Urbana-Champaign, Champaign, IL

This study will investigate using response times (RTs) with item responses in CAT to enhance item selection and ability estimation. Using van der Linden's (2007) hierarchical framework, a maximum likelihood procedure for joint estimation of ability and speed parameters along with information functions for use in CAT are developed.

Maximizing Information Per Unit Time in Adaptive Testing

Ying Cheng, University of Notre Dame, Notre Dame, IN and John Behrens, Pearson, Mishawaka, IN

Does selecting items with maximum information per unit time automatically control item exposure? Preliminary results show that corr(item discrimination, log(response time)) = .21, suggesting that more discriminating items take longer to finish. We plan to conduct a more thorough real data analysis and simulation study to answer the main question.

Sunday, April 6, 2014 • 2:30 PM - 4:10 PM, Regency C1
Paper Session, M4

Diagnostic and Formative Assessments: Modeling Strategies

Session Chair: Hidetoshi Saito, Ibaraki University, Mito, Japan

Cognitive Diagnostic Test Model of Pre-Literacy Implements Semiotics and DCT

Eilene Edejer, Loyola University, Chicago, IL and Nikolaus Bezruczko, Measurement and Evaluation Consulting, Chicago, IL

Cognition research and semiotics were conceptually integrated to establish a diagnostic pre-literacy test model. Partial credit analysis of authentic child samples was coherent and reliable. Decomposition of item difficulties into sensorimotor, iconic, and symbolic components accounted for 70 percent of variance. Validation supported test model results and preschool outcomes.

Using the LEAFF Model to Enhance Classroom Learning and Assessment

Jacqueline P. Leighton, Gomez Bustos, and Maria Clara, CRAME/University of Alberta, Edmonton, Canada

The LEAFF model indicates causal connections between cognitive and emotional variables in students' learning and assessment. An experimental study was conducted. Results indicated that an intervention led to stronger feelings of student comfort during instruction and greater ease indicating material that was confusing. Implications for formative assessment feedback are discussed.

The Use of Formative Assessment in Japanese Middle School Classrooms

Hidetoshi Saito, Ibaraki University, Mito, Japan

The present study is a mixed method study of the current formative assessment (FA) practice of junior and senior high school EFL teachers in Japan. Survey data from 732 teachers and qualitative data from four of those teachers identify three types of FA users and reveal common practices of FA.

Development and Construct Validity of the Classroom Strategies Scales-Teacher Form

Linda Reddy, Stephanie Peters, and Christopher Dudek, Rutgers University, New Brunswick, NJ

This article presents the validation of a teacher observational measure, Classroom Strategies Scale-Teacher Form. Confirmatory factor analyses of CSS data from 317 classrooms tested hypothesized four-, five- and six-factor models. The CSS evidences overall good reliability indices and construct validity. Implications for the CSS in educator evaluation will be discussed.

Sunday, April 6, 2014 • 2:30 PM - 4:10 PM, Regency C2
Paper Session, M5

Single Item Scale/Computer & Paper/Jackknifed Variance Estimation

Session Chair: Megan Welsh, University of Connecticut, Storrs, CT

Understanding Differential Item Functioning With the General Linear Model

Jessica Loughran and William Skorupski, University of Kansas, Lawrence, KS

This study used the general linear model (GLM) to explain differential item functioning (DIF) statistics from specific item characteristics. DIF statistics for English language learners (ELLs) versus non-ELLs on a math test were obtained using multiple DIF methods. Results show that certain linguistic item characteristics significantly explain DIF.

Reviewing the Literature on Linking Paper and CBT Scores

Jinghua Liu, Secondary School Admission Test Board, Skillman, NJ and Neil J. Dorans, ETS, Princeton, NJ

When test forms are administered in different modes, e.g., paper and computer, it is necessary to determine what interpretation to attach to linkages between these scores. This paper reviews previous studies that examined score linkages between computer-based and paper-pencil tests.

Robust Grouped Jackknifed Variance Estimation

Jiahe Qian and Shelby Haberman, Educational Testing Service, Princeton, NJ

Variance estimation by jackknifing is quite sensitive to outliers. Alternative variance estimates are considered based on trimmed and Winsorized pseudo-values. Results are illustrated with real survey data.

Examining Bias on a Single Item Scale With Think Alouds

Megan Welsh, Janice Kooken, Faith G. Miller, Rivah Rosen, Sandra M. Chafouleas, University of Connecticut, Storrs, CT; Gregory A. Fabiano, SUNY Buffalo, Buffalo, NY; and T. Chris Riley-Tillman, University of Missouri, Columbia, MO

The study investigates test bias in a single-item behavior rating scale using verbal protocol analysis. Teachers reviewed and scored videos of lessons in which the behavior of children of varying races/sexes was controlled so that biased inferences could be identified. Results and potential applications to other instruments are discussed.

Sunday, April 6, 2014 • 2:30 PM - 4:10 PM, Commonwealth A Coordinated Session, M6

Theories of Action About Performance Assessment: Assessment Consortia Validation Research

Organizer and Session Chair: Steve Ferrara, Pearson, Washington, DC

Validity Evaluation in the NCSC Alternate Assessment Context

Ellen Forte, edCount, Washington, DC

The NCSC alternate assessment consortium is using an argument-based approach to guide its validity evaluation and targeting validity issues that are, in some cases, unique to the alternate assessment context and require new or revised methods to gather evidence. This paper focuses on these unique design and evaluation challenges.

Theories of Action About Performance Assessments: The ASSETS Consortium

Howard G. Cook, University of Wisconsin, Oregon, WI

This presentation describes the components of the ASSET English Language Proficiency Assessment Consortium system, especially as they relate to performance-based item formats, and discusses the Consortium's theory-of-action, current research objectives and findings in support of its theory-of-action claims.

Multi-State Consortium Performance Assessment: Theory into Action

Martha McCall, Smarter Balanced Assessment Consortium, Portland, OR

Results of a multi-state consortium pilot test are discussed in this paper. Studies focus on association patterns shown by performance task responses, examining coherence of performance task items within the task and degree of dimensionality with other test items. Student engagement and performance are also reported.

Design and Development of PARCC Performance-Based Assessments and Related Research

Enis Dogan, PARCC, Washington, DC

In this paper we illustrate sample PARCC performance-based tasks and discuss the principles followed in the design and development of these tasks. In addition, we also discuss specific studies undertaken in spring and summer of 2013 to guide development of ELA/L and mathematics PB tasks.

Sunday, April 6, 2014 • 2:30 PM - 4:10 PM, Commonwealth B
Paper Session, M7

Cut-Scores, Subsets and Smoothing in Standard Setting

Session Chair: Brett P. Foley, Alpine Testing Solutions, Denton, NE

Evaluating the Operational Feasibility of Using Item Subsets for Standard-Setting

Priya Kannan, Adrienne Sgammato, Richard J. Tannenbaum, and Irvin R. Katz, Educational Testing Service, Princeton, NJ

Building on results from previous resampling studies, in this study, we evaluated the feasibility of using item subsets in an operational setting. Cut-scores recommended by a panel using 45 items (a number suggested by previous work) were found to be comparable to those recommended by another panel using the entire test.

Choosing Cut-offs for Related Tests

Dan B. Wright, ACT Inc., Austin, TX

Six procedures for estimating “on target” cut-offs for educational tests are compared. The choice of which procedure to use can make a difference. Overall, a method based on a loess regression appears the most promising of those compared. Recommendations for estimating and reporting cut-offs are made.

Evaluating an Impact Percentage Smoothing Vertically Moderated Standard Setting Design

Brett P. Foley, Alpine Testing Solutions, Denton, NE

This study provides an illustration of a new VMSS design incorporating multiple approaches to help ensure coherence of performance expectations across grade levels. The study illustrates a successful combination of several VMSS features recommended in the literature and expands the scope of VM SS research to students with disabilities.

Sampling Distributions of Cutscores Based on Panelist Accuracy and Consistency

William P. Skorupski and Joseph Fitzpatrick, University of Kansas, Lawrence, KS

The purpose of this study was to evaluate the effects of panelist accuracy and consistency on resulting cutscore recommendations. A simulation was conducted to create plausible Angoff standard setting data to determine how the sampling distribution of cutscores is affected by changes in these factors.

Sunday, April 6, 2014 • 2:30 PM - 4:10 PM, Commonwealth C
Paper Session, M8

DIF Analysis Methods

Session Chair: Rebecca Zwick, Educational Testing Service, Santa Barbara, CA

Extending the Empirical Bayes DIF Procedure to Polytomous Items

Rebecca Zwick, Educational Testing Service, Santa Barbara, CA; Lei Ye, and Steven Isham, Educational Testing Service, Princeton, NJ,

The empirical Bayes DIF method of Zwick, Thayer, and Lewis has been found to yield more accurate DIF results than standard approaches. We have extended the model to polytomous items. We will report results of a simulation that evaluates both decision accuracy and the properties of the DIF estimates.

Crossing SIBTEST in a Multilevel Data Environment

Brian French, Washington State University, Pullman, WA and Holmes Finch, Ball State University, Muncie, IN

This simulation study examines the performance of non-uniform differential item functioning (DIF) detection techniques in a multilevel data environment. Evaluation occurred for standard crossing-SIBTEST and three proposed extensions of crossing-SIBTEST designed to account for multilevel data. Differences in DIF detection are noted when the analytic strategy matches the data structure.

Nonparametric Regression for DIF Detection

Qiwen Zheng, University of Maryland, College Park, MD; Shanghong Xie, and Jeff Douglas, UIUC, Champaign, IL

Isotonic regression provides a method for fitting regression functions nonparametrically that satisfy a monotonicity constraint. This is appealing for item characteristic curves, and the monotonicity constraint is easily enforced through the pooled adjacent violators algorithm that results in a least squares estimator subject to the monotonicity constraint. Here we develop a statistical test for DIF comparing ICCs of two samples. An R package for the procedure is also introduced.

Detecting Differential Statement Functioning in Ipsative Tests Using the Logistic Regression Method

Chia-Wen Chen and Wen-Chung Wang, The Hong Kong Institute of Education, Hong Kong, Hong Kong

The Rasch ipsative model has developed for ipsative data, in which each statement has a utility. In practice, a statement may exhibit different utilities for different respondents. Using simulations, we evaluated the logistic regression method in detecting the differential statement functioning. The results supported its feasibility.

Sunday, April 6, 2014 • 2:30 PM - 4:10 PM, Commonwealth D
Paper Session, M9

Subscores and Error Information

Session Chair: Christine Hutchison, University of Virginia, Lynchburg, VA

Comparing Procedures of Subscale Score Reporting in CAT

Changjiang Wang, Chingwei D. Shin, and Yuehmei Chien, Pearson, Iowa City, IA

Subscore reporting has been a topic that attracts much research effort and different procedures have been proposed. Based on the review of these procedures, this study evaluates the precision of subscale score estimation under different simulated conditions in a computer adaptive setting.

Validating a Mathematics Interim Assessment With Cognitively Diagnostic Error Categories

Christine Hutchison, University of Virginia, Lynchburg, VA and Patrick Meyer, University of Virginia, Charlottesville, VA

The purpose of this mixed methods study is to create and validate a multiple-choice mathematics interim assessment where the distractors are cognitively diagnostic. Validity evidence will comprise polytomous item analyses (e.g., CTT, DIF, and IRT methods), descriptive statistics of cognitive think-alouds, and a Chi-square analysis of student error categories.

Error Feature Extractions From Algebra Templates

Thomas S. McTavish, Pearson, Denver, CO and Johann A. Larusson, Pearson, Boston, MA

With only final answers submitted to algebra exercises in a homework system that uses templates, this paper demonstrates methods of enumerating different types of mechanical errors and misconceptions. When set against the parameter space of correct solutions, the distribution of errors can reveal intra-template variability and enable higher resolution assessments.

Using Likelihood Ratio Tests to Evaluate the Uniqueness of Subscores

Peter Baldwin, National Board of Medical Examiners, Philadelphia, PA

In addition to overall scores, examiners frequently elect to report subscores corresponding to major content subareas; however, this activity does not guarantee that subscores measure distinct traits. This paper describes how likelihood ratio tests can be used to measure the uniqueness of subscores compared to scores for other item subsets.

Sunday, April 6, 2014 • 4:00 PM–7:00 PM
Jefferson Boardroom

NCME Board of Directors Meeting

Members of NCME are invited to attend as observers.

Sunday, April 6, 2014 • 4:20 PM - 6:00 PM, Washington
Coordinated Session, N1

English-Learner Measurement: Score Interpretation, Performance Standards and Redesignation Impact

Organizer and Session Chair: Molly M. Faulkner-Bond, University of Massachuetts at Amherst, Northampton, MA

Academic Language and Academic Performance: A Multilevel Study of ELs

Molly M. Faulkner-Bond, University of Massachusetts at Amherst, Northampton, MA and Mikyung K. Wolf, Educational Testing Service, Princeton, NJ

We present the results of a multi-level model examining the relationship between academic language (AL) and content test performance for fourth grade English learners nested within schools in three states. We describe student- and school-level results, including considerations about which school-level features might have fixed effects on the language-content relationship.

Validity and Effectiveness of Accommodations for ELLs

Jamal Abedi, University of California at Davis, Davis, CA

We present results from a study evaluating the effectiveness and validity of two accommodations for English learners (ELs): English glossary and customized English dictionary. Both led to significant performance increases for ELs, and had no effect on the performance of non-ELs, supporting the hypothesis that both accommodations are effective and valid.

What's "English Proficient?" How Long Does it Take to Attain?

Gary Cook, Wisconsin Center for Education Research, Madison, WI and Robert Linquanti, WestEd, Oakland, CA

This presentation highlights results from a study that examines 1) how English proficiency might be determined empirically and 2) how long it takes for English learners (ELs) to become English proficient. The methods explored can support states in establishing an English proficient criterion for ELs.

Classification Models and English Learner Redesignation: High-Performing Students Left Behind?

Patricia Carroll and Alison Bailey, University of California at Los Angeles, Los Angeles, CA

English Language Proficiency Assessments (ELPAs) assess the academic language skills of language minority students to determine language programming. This descriptive study applies six classification models to actual ELPA performances K-5 to estimate model impact on redesignation eligibility. Findings suggest models differ in non-proficient and redesignation-eligible percentages, even among high-performing students.

Sunday, April 6, 2014 • 4:20 PM - 6:00 PM, Regency A
Coordinated Session, N2

Using Enhanced CAT Designs to Improve Future Implementations

Organizer and Session Chair: Haiyan Lin, ACT, Inc., Iowa City, IA

Using Collateral Information in a Computerized Adaptive Test Battery

Xinrui Wang, Pearson VUE, Chicago, IL, Yuki Nozawa, Benesse Corporation, Tokyo, Japan; Xiaohong Gao and Haiyan Lin, ACT Inc., Iowa City, IA

This study investigates an optimal adaptive strategy for a large-scale classification test battery involving three subtests to improve ability estimation accuracy and test efficiency. Three factors are explored: (1) order of subtests, (2) different entry points for estimation reflecting different levels of utilizing collateral information, and (3) estimation methods.

Multiple-Pool Design in CAT: Is it Necessary or Not?

Chanjin Zheng, UIUC, Champaign, IL, Shiyu Wang, University of Illinois at Urbana-Champaign, Urbana, IL, and Chunyan Liu, ACT Inc., Iowa City, IA

This study investigates the effectiveness of multiple-pool design for enhancing test security when it is used alone or combined with item exposure control methods. The preliminary result indicates that the multiple-pool design demonstrates little effect on item usage if effective item exposure control is used.

Hybrid Designs for Computer-Based Adaptive Testing

Shiyu Wang, University of Illinois at Urbana-Champaign, Urbana, IL and Haiyan Lin, ACT, Inc., Iowa, IA

This study proposes general rules for hybrid designs, which combine CAT and MST. Generalizations of the rules to three specific simulated test situations were illustrated as examples. Simulation studies were conducted to verify the rules, expected properties, and advantages of these newly proposed designs

Modified Block Review Approaches to Allow Answer Change in CAT

Haiyan Lin, ACT, Inc., Iowa City, IA

This study investigates procedures for improving ability estimation accuracy in reviewable CAT. Three proposed approaches used a-stratification to modify the block review method, allowing examinee answer changes. The fisher information, Kullback-Leibler information, and Shannon-entropy method were used for item selection. The Item Pocket method was explored and compared too.

Sunday, April 6, 2014 • 4:20 PM - 6:00 PM, Regency B
Coordinated Session, N3

Automatic Scoring of Non-Traditional Forms of Assessment

Organizer and Session Chair: F. Jay Breyer, ETS, Princeton, NJ

Multivariate Outliers in Automated Essay Scoring

Jiyun Zu, Isaac Bejar, and Matthew Duchnowski, Educational Testing Service, Princeton, NJ

The study explores ways to improve automated scoring of essays based on regression diagnostics. We identify essays that are atypical in the predictor space as well as cases that have large residuals. We then conduct expert analysis of the essays to identify the reason behind atypicality and misfit.

Weighting Human and Machine Scores to Compensate for Construct Misalignment

F. Jay Breyer and Florian Lorenz, ETS, Princeton, NJ

Written tasks that require demonstration of comprehension as well as writing proficiency present unique challenges to aligning adequate construct measurement with known shortcomings of automated scoring. A differential weighting method that combines human and automated scores in equal and unequal proportions is evaluated for construct fidelity as the reported score.

Combination of Scores With the Automated Score Monitoring System

Su-Youn Yoon, Educational Testing Service, Princeton, NJ

This study proposes an efficient method to combine human and automated scoring. An automated score monitoring system identifies the responses for which automated scores are likely to be incorrect and routs the responses to human raters. This approach increases scoring accuracy with a small amount of the human scoring.

Practical Effects of Different Read-Behind Scenarios in Automated Essay Scoring

Vincent Kieftenbeld and Michelle Barrett, CTB/McGraw-Hill Education, Monterey, CA

This study seeks to determine the practical effect of using different engine/human scoring scenarios on student scores of record. The study will have relevance for users of automated scoring who are considering the cost and score implications of using automated scoring under a variety of scenarios.

Sunday, April 6, 2014 • 4:20 PM - 6:00 PM, Regency C1
Paper Session, N4

Observation Instruments and Rating

Session Chair: Carrie Semmelroth, Boise State University, Boise, ID

Validating an Observation Tool to Measure Special Education Teacher Effectiveness

Carrie Semmelroth and Evelyn Johnson, Boise State University, Boise, ID

This presentation uses Kane's (2006) validity argument approach to validate an observation tool that evaluates special education teacher effectiveness through the implementation of evidence-based instructional practices. Relevant evidence for two of the four validity inferences (scoring and generalization) will be interpreted and applied towards the development of the special education observation tool.

Validity Evidence of Value-Added Model Scores Using an Observational Rubric

Claudia Guerere, University of South Florida, Largo, FL and Robert F. Dedrick, University of South Florida, Tampa, FL

Value-added model scores are used for high-stakes decision making. Validity evidence of these scores is critically needed. This research, using data from 2385 teachers from 104 schools in one school district in Florida, examined the validity of the value-added scores by correlating these scores with scores from an observational rubric.

The Effect of Raters and Schools on Teacher Observation Scores

Jianlin Hou, The School District of Palm Beach, Wellington, FL; Bob Johnson, Bidya Subedi, and Mark Howard, The School District of Palm Beach, Palm Beach, FL

Differences between raters and schools have large effects on the teacher observation scores used in teacher evaluations. This study uses multilevel model to quantify the proportion of variance attributable to raters, schools, and individual teachers in a large urban school district for two consecutive years.

Rating Design and the Reliability and Validity of Teaching Observations

Patrick Meyer, University of Virginia, Charlottesville, VA, Andrew J. Mashburn, Portland State University, Portland, OR; Joseph P. Allen, and Robert C. Pianta, University of Virginia, Charlottesville, VA

We conducted an experimental study that manipulated observation length and presentation order of videotaped lessons from secondary classrooms. Randomly presented segments produce the most desirable effect on score reliability and validity. Results also suggest that randomizing observation order reduces construct irrelevant variance arising from carry over effects and rater drift.

Sunday, April 6, 2014 • 4:20 PM - 6:00 PM, Regency C2
Paper Session, N5

Accuracy and Reliability of Diagnostic Information and Raters

Session Chair: Chi Chang, Michigan State University, Okemos, MI

Estimating Attribute-Level and Pattern-Level Classification Consistency and Accuracy Indices

Wenyi Wang, Lihong Song, Shuliang Ding, Jiangxi Normal University, Nanchang, China; and Jiayuan Yu, Nanjing Normal University, Nanjing, China

Cognitive diagnostic assessment focuses on make attribute classification-based decisions while reliability of classification is often not reported in diagnostic score reporting. The study investigates attribute-level and pattern-level classification consistency and accuracy indices based on the Rudner method or the Guo method. Then we use simulated data to evaluates their performances.

The Effect of Termination Criteria on Cognitive Diagnostic Computerized Adaptive Testing

Chi Chang, Michigan State University, Okemos, MI and Liyang Mao, Michigan State University, East Lansing, MI

The purpose of this study is to investigate the accuracy, efficiency and consistency of three stopping rules in CD-CAT: SPRT, CI, and Tatsuoka's approach (2002). Cognitive diagnostic assessment elements were embedded in each components of CAT and the performances of 36 conditions are going to be compared in the study.

A Caution in the Quest for Diagnostic Test-Based Inferences

Laine Bradshaw, The University of Georgia, Athens, GA and James Koepfler, Pearson, Wilmington, DE

To meet the high demands for educational tests that provide diagnostic feedback, researchers have proposed methods for extracting multidimensional diagnoses from unidimensional test results. This study demonstrates why these methods can yield invalid inferences and illustrates how using this feedback may misguide instructional efforts meant to improve student achievement.

Using Inferential Confidence Intervals as a Criterion for Rater Consistency

Frank E. Williams, Educational Testing Service, Princeton, NJ

This paper provides an alternative to subtracting agreement statistics for assessing similarity between raters by incorporating the sampling distribution of the agreement statistics. Using confidence intervals is a conservative approach but may be more suitable for high-stakes tests such as the writing test used in this study.

Sunday, April 6, 2014 • 4:20 PM - 6:00 PM, Commonwealth A
Coordinated Session, N6

Practical Approaches to Defining the Test Construct and Specifications

Organizer and Session Chair: Kirk A. Becker, Pearson, Chicago, IL

Experimental Comparison of Content-Matched Multimedia and Traditional Formats

Kathleen Z. Holtzman and David B. Swanson, NBME, Philadelphia, PA

This presentation describes two studies of incorporating multimedia into MCQ-based medical licensure and certification exams. For both, empirical comparisons of the measurement characteristics of multimedia-based test material with content-matched control material were conducted to quantify trade-offs between greater authenticity in presentation of findings and increased testing time requirements.

De-Constructing Constructs: Evaluating Stability of Higher-Order Thinking Across Technology-Rich Scenarios

Lisa A. Keller, University of Massachusetts at Amherst, Amherst, MA, April L. Zenisky, University of Massachusetts at Amherst, South Hadley, MA, and Xi Wang, University of Massachusetts at Amherst, Amherst, MA

With various emerging measurement formats increasingly considered for large-scale tests, measurement analysis and assumptions must be reexamined. The present study examines the comparability of task analysis strategies across two technology-rich scenarios to evaluate the construct consistency across scenarios and the implications of scenario use for data analysis in operational testing.

Defining the Innovative Item: Test Constructs and Item Prototypes

Kirk A. Becker, Pearson, Chicago, IL and Belinda Brunner, Pearson, Manchester, United Kingdom

Well-crafted assessments are the result of blending technology with sound test design. This presentation provides a guided process for conducting a multidisciplinary Content Design Workshop to with the purpose of designing item prototypes for testing subjects such as higher-level thinking skills.

Developing a Practice Analysis (PA) for a Skill-Based Profession

Joshua Stopek, AICPA, Ewing, NJ, Matthew J. Burke, AICPA, Yardley, PA, and Henrietta Eve, AICPA, Ewing, NJ

This presentation describes the work being done to develop a PA survey for a Licensure examination in a professional field. It will focus on how principled assessment frameworks are helpful in providing the infrastructure for representing the information necessary to thoughtfully consider all aspects of the construct to be measured.

Sunday, April 6, 2014 • 4:20 PM - 6:00 PM, Commonwealth B Coordinated Session, N7

Strengthening Student Assessment in Developing Countries

Organizer and Session Chair: Marguerite Clarke, World Bank, Washington, DC

Overview of the SABER-Student Assessment and READ Programs

Marguerite Clarke, World Bank, Washington, DC

This presentation describes the conceptual framework and diagnostic tools developed under the SABER–Student Assessment and READ programs. The tools are used by countries to benchmark their student assessment systems and to make strategic decisions about where to introduce reforms.

Benchmarking Student Assessment Systems in Developing Countries

Julia Liberman, World Bank, Washington, DC

The objective of this presentation is to describe the results obtained from benchmarking the eight READ countries' student assessment systems using the diagnostic tools based on the SABER-Student Assessment framework.

Supporting the Development of Student Assessment Systems in READ Countries

Emily Gardner, World Bank, Washington, DC

This presentation discusses the implementation of the strategies used by Armenia, Mozambique, and Tajikistan to strengthen their student assessment systems. The benchmarking results were used as input to a variety of reforms, albeit all had the same goal of producing better assessments that would support better student learning.

Reflective Contribution By External Discussant

Alan Ruby, University of Pennsylvania, Philadelphia, PA

Sunday, April 6, 2014 • 4:20 PM - 6:00 PM, Commonwealth C
Paper Session, N8

Special DIF Analysis

Session Chair: Kevin Cappaert, University of Wisconsin at Milwaukee, Milwaukee, WI

Assessing DIF in Multidimensional Items With the Logistic Regression Method

Hui-Fang Chen, City University of Hong Kong, Hong Kong; Kuan-Yu Jin and Wen-Chung Wang, Hong Kong Institute of Education, Hong Kong

DIF can occur in multidimensional items. When the logistic regression method is adopted for DIF assessment in multidimensional items, the selection of matching variables becomes a critical issue. The simulation results suggest the selection should match the dimensionality of the studied item to yield high power of DIF detection.

Detection of Differential Item Functioning for Repeated Measures

Jennifer Lord-Bessen, Ying Liu, and Arnond Sakworawich, Fordham University, Bronx, NY

This paper proposes an innovative procedure, based on alternative logistic regression, to detect DIF in repeated measures. Simulations show that the proposed methodology is effective and powerful, providing a valuable tool in removing instrumentation effect from longitudinal studies and ensuring the measurement of true change in trait is not contaminated.

Simultaneous Uniform and Non-Uniform DBF Detection: A MIMIC Model Approach

Kevin Cappaert, University of Wisconsin at Milwaukee, Milwaukee, WI, Holmes Finch, Ball State University, Muncie, IN, and Cindy M. Walker, University of Wisconsin at Milwaukee, Milwaukee, WI

The multiple indicator multiple cause (MIMIC) model has shown to be a viable alternative to simultaneous item bias test (SIBTEST) for uniform DBF detection, especially in instances of impact. A proposed extension of the MIMIC model to include non-uniform DBF detection will be investigated in relation to crossing SIBTEST.

Differential Item Functioning in a Multi-Stage Adaptive Testing Environment

Kelvin D. Gregory and Goran Lazendic, Australian Curriculum, Assessment and Reporting Authority, Sydney, Australia

This study evaluated methods to identify differential item functioning in a multistage computer adaptive test. The power and type 1 error rates for Mantel-Haenszel and logistic regression techniques were investigated using simulations approximating two conditions encountered in sample-based Australian testing. The impact of DIF on group ability estimates is reported.

Sunday, April 6, 2014 • 4:20 PM - 6:00 PM, Commonwealth D
Paper Session, N9

Diagnostic Assessment and Classification

Session Chair: Fei Zao, Arizona Department of Education, Phoenix, AZ

Accuracy of Neural Network versus Nonparametric Approaches in Diagnostic Classification
Thomas P. McCoy and John T. Willse, UNC Greensboro, Greensboro, NC

Diagnostic classification models (DCMs) are growing in popularity, but large samples are needed for item calibration. Neural network and nonparametric approaches have been proposed for DCM classification, and could be promising for providing formative feedback in desired educational settings. A simulation study on classification rates is presented.

Bayesian Evaluation of Informative Hypothesis in Diagnostic Assessment of Mathematics
Jorine Vermeulen, University of Twente, Tilburg, Netherlands, Herbert Hoijtink, Utrecht University, Cito Institute for Educational Measurement, Utrecht and Theo Eggen, University of Twente, Cito Institute for Educational Measurement, Enschede

The value of cognitive diagnostic assessment for formative use is increasingly being recognized. The empty number line was used to collect data about the solution process of grade 3 students, and their conceptual understanding of subtraction and addition was diagnosed by using Bayesian evaluation of informative diagnostic hypotheses.

Measurement Error in Latent Variables That Predict STEM Retention
Lynne S. Schofield, Swarthmore College, Swarthmore, PA

This paper investigates models that attend to measurement error bias when using latent variables as predictors in regression analyses. A structural equations model incorporating item response theory is proposed to overcome the shortcomings of current methods. Using the new model, I examine racial and gender differentials in STEM retention.

Cognitive Diagnostic Model Comparisons Using Empirical Data
Fei Zhao, Arizona Department of Education, Phoenix, AZ and Neal Kingston, University of Kansas, Lawrence, KS

Three cognitive diagnostic models are compared using empirical data. Parameter estimation, global model-fit (using PPMC), and model reliability information are provided. Models are further compared on different examinee mastery levels on each of the skills. The examinee classification results for each of the model pairs are also presented.

Participant Index

Participant Index

Participant Index

Participant Index

Participant Index

Participant Index

Participant Index

Participant Index

Participant Index

Participant Index

Participant Index

Contact Information for Individual and Coordinated Sessions First Authors

Abedi, Jamal, University of California, Davis, Davis, California

Ainley, John, ACER, 19 Prospect Hill Road, Camberwell, Victoria, 3124, Australia, john.ainley@acer.edu.au

Akbay, Lokman, Rutgers, The State University of New Jersey, 46 Marvin Lane, Piscataway, New Jersey, 08854, lokmanakbay@gmail.com

Aksu Dunya, Beyza, University of Illinois at Chicago, 1040 W Harrison Street, Chicago, Illinois, 60607, baksu2@uic.edu

Albanese, Mark A., National Conference of Bar Examiners, 302 South Bedford Street, Madison, Wisconsin, 53705, maalbane@wisc.edu

Albano, Anthony, University of Nebraska, Lincoln, albano@unl.edu

Ali, Usama, ETS, Princeton, New Jersey, uali@ets.org

Allalouf, Avi, NITE, Jerusalem, Israel, avi@nite.org.il

Almehrizi, Rashid S., Sultan Qaboos University, Muscat, Oman, mehrzi@squ.edu.om

Almond, Russell, Florida State University, 1114 W. Call Street, Tallahassee, Florida, 32306, ralmond@fsu.edu

Almonte, Debby E., Educational Testing Service, Rosedale Road, Princeton, New Jersey, 08541, dalmonte@ets.org

Anderson, Carolyn J., University of Illinois, 1310 South Sixth St., MC-708, Champaign, Illinois, 61820, cja@illinois.edu

Anderson, Margaret, University of Kansas, 309 Elm St., Lawrence, Kansas, 66044, ganders@ku.edu

Andrews, Benjamin, ACT, Iowa City, Iowa, benjamin.andrews@act.org

Andrich, David, The University of Western Australia, Graduate School of Education, 35 Stirling Highway, Crawley, Western Australia, 6009, Australia, david.andrich@uwa.edu.au

Arce-Ferrer, Alvaro J., Pearson, San Antonio, Texas, 78258, alvaro.arce-ferrer@pearson.com

Arenson, Ethan, National Board of Osteopathic Medical Examiners, 8765 W. Higgins Rd, Suite 200, Chicago, Illinois, 60631

Arieli-Attali, Meirav, Educational Testing Service, Princeton, New Jersey, mattali@ets.org

Atalmis, Erkan, bereket@ku.edu

Attali, Yigal, ETS, Princeton, yattali@ets.org

Bailey, Katharine, CEM, Durham University, Mountjoy Rowan Block, Stockton Road, Durham, DH1 3UZ, United Kingdom, kate.bailey@cem.dur.ac.uk

Baldwin, Peter, National Board of Medical Examiners, 3750 Market Street, Philadelphia, Pennsylvania, 19104, pbaldwin@nbme.org

Bates, Michael, Michigan State University, East Lansing, Michigan

Baudet, Alexandre, CRP Henri Tudor, Luxembourg, Luxembourg

Bay, Luz, College Board, 290 Long Hill Rd, Dover, New Hampshire, 03820, lbay@collegeboard.org

Beaver, Jessica L., Washington State University, Cleveland 80, Pullman, Washington, 99164, jessica.l.beaver@email.wsu.edu

Becker, Kirk A., Pearson, 2415 N Drake, Chicago, Illinois, 60647, kirk.becker@pearson.com

Béguin, Anton A., Cito Institute for Educational Measurement, PO Box 1034, Arnhem, 6801 MG, Netherlands, anton.beguin@cito.nl

Behrens, John T., Pearson, Center for Digital Data, Analytics & Adaptive Learning, 15300 Fox Run Trail, Mishwaka, Indiana, 46545, john.behrens@pearson.com

Bejar, Isaac, ETS, 32 Arrowwood Dr., Hamilton, New Jersey, 08690, ibejar@ets.org

Belov, Dmitry, Law School Admission Council, Newtown, Pennsylvania, dbelov@lsac.org

Bennett, Randy, ETS, Princeton, New Jersey, 08541, rbennett@ets.org

Bergner, Yoav, Educational Testing Service, Princeton, ybergner@ets.org

Contact Information for Individual and Coordinated Sessions First Authors

Bertling, Jonas P., Educational Testing Service, Princeton, New Jersey, jbertling@ets.org

Betts, Joseph, Houghton Mifflin Harcourt, 710 W Burr Oak Dr., Arlington Heights, Illinois, 60004, jbetts5118@aol.com

Bezruczko, Nikolaus, Measurement and Evaluation Consulting, 1524 E. 59th Street, A-1, Chicago, Illinois, 60637, nbezruczko@msn.com

Bo, Yuanchao, Fordham University, 441 E Fordham Rd, Dealy Hall 0226 Psych, Bronx, New York, 10458, boyuanchao@gmail.com

Bolsinova, Maria, Utrecht University / Cito Institute for Educational Measurement, Utrecht, Netherlands, m.a.bolsinova@uu.nl

Boyd, Aimee, Pearson, 400 Center Ridge Drive, Austin, Texas, aimee.boyd@pearson.com

Bradshaw, Laine, The University of Georgia, 323 Aderhold Hall, Athens, Georgia, laineb@uga.edu

Braslow, David, Harvard Graduate School of Education, Cambridge

Braun, Henry, Boston College, Boston

Breyer, F. Jay, ETS, Rosedale Raod, Mail Stop 18-E, Princeton, New Jersey, 08641, fbreyer@ets.org

Briggs, Derek, University of Colorado, 249 UCB, Boulder, Colorado, 80309, derek.briggs@colorado.edu

Broaddus, Angela, University of Kansas, Lawrence, Kansas

Brown, Christy, Clemson University, Clemson, South Carolina, cjb2@clemson.edu

Brown, Nathaniel J. S., Boston College, Chestnut Hill

Buchholz, Janine, DIPF (German Institute for International Educational Research), Unterlindau 43, Frankfurt, 60323, Germany, buchholz@dipf.de

Buckley, Katie H., Harvard University, Cambridge, buckley.kate@gmail.com

Bulut, Okan, American Institutes for Research, Washington, District of Columbia

Burke, Matthew J., AICPA, 215 Taylorsville Road, Yardley, Pennsylvania, 19067, mburke@aicpa.org

Burstein, Jill, ETS, Princeton, New Jersey, 08540, jburstein@ets.org

Buzick, Heather M., Educational Testing Service, 660 Rosedale Road, MS 10R, Princeton, New Jersey, 08541, hbuzick@ets.org

Camara, Wayne, ACT, Princeton

Cao, Yi, ETS, Princeton, ycao@ets.org

Cappaert, Kevin, University of Wisconsin - Milwaukee, Milwaukee, Wisconsin, 53211, cappaer3@uwm.edu

Carroll, Patricia, University of California, Los Angeles, Los Angeles, California

Casabianca, Jodi M., University of Texas - Austin, Austin, Texas, jcasabianca@austin.utexas.edu

Castellano, Katherine E., University of California, Berkeley, Berkeley, California

Chajewski, Michael, The College Board, 45 Columbus Avenue, New York, New York, 10023, mchajewski@collegeboard.org

Chang, Chi, Michigan State University, 4800 Country Way East Apt. 304, Okemos, Michigan, 48864, changc65@msu.edu

Chang, Yu-Feng, University of Minnesota, Minneapolis, Minnesota

Chao, Hsiu-Yi, National Chung Cheng University, Dept. of Psychology, National Chung Cheng University, 168, Sec. 1, University Rd., Minhsiung Township, Chiayi County, 62102, Taiwan, hsiuyi1118@gmail.com

Chapman, Allison E. A., Queen's University, Kingston, Canada, chapman.a@queensu.ca

Chen, Chia-Wen, The Hong Kong Institute of Education, Hong Kong, Hong Kong, five43@gmail.com

Chen, Cong, University of Illinois at Urbana-Champaign, 1310 S. Sixth Street, Champaign, Illinois, 61820, cchen105@illinois.edu

Chen, Feng, University of Kansas, Lawrence, Kansas, 66045

Contact Information for Individual and Coordinated Sessions First Authors

Chen, Haiqin, University of Missouri, Columbia, hc9pd@mail.missouri.edu

Chen, Hui-Fang, City University of Hong Kong, Hong Kong, Hong Kong, g8932006@gmail.com

Chen, Jie, ACT, Inc., Iowa City, xiaojiewd@hotmail.com

Chen, Juan, University of Iowa, Iowa City, Iowa, juan-chen-1@uiowa.edu

Chen, Ping, Beijing Normal University, Room 122, National Assessment of Education Quality, No.19, Xinjiekouwai Street, Haidian District, Beijing, 100875, China, pchen@bnu.edu.cn

Chen, Yi-Hsin, University of South Florida, Tampa, ychen5@usf.edu

Cheng, Yi-Ling, Michigan State University, 1547 N. Hagadorn Rd. Apt 29, East Lansing, Michigan, 48823, chengyil@msu.edu

Cheng, Ying, University of Notre Dame, Notre Dame, Indiana, 46530, ycheng4@nd.edu

Chien, Yuehmei, Pearson, 1260 Tipperary Rd, Iowa City, Iowa, 52246, yuehmei.chien@pearson.com

Chin, Mark, Harvard University, Center for Education Policy Research, 50 Church Street, 4th Floor, Cambridge, Massachusetts, 02138, mark_chin@gse.harvard.edu

Chinn, Roberta N., 113 Farham Drive, Folsom, California, 91505, r.chinn@comcast.net

Cho, Sun-Joo, Vanderbilt University, 700 12 Ave S., Nashville, Tennessee, 37203, sjcho0306@gmail.com

Choe, Edison, University of Illinois Urbana-Champaign, Champaign, Illinois, 61820, emchoe2@illinois.edu

Choi, In-Hee, University of California Berkeley, Graduate School of Education, 685 Liberty Ship Way #300, Albany, California, 94706, ineechoi@berkeley.edu

Choi, Jinah, University of Iowa, Iowa City, jinah-choi@uiowa.edu

Choi, Jinnie, Rutgers, The State University of New Jersey, New Brunswick, jinnie.choi@gmail.com

Choi, Youn-Jeng, University of Georgia, Athens, neatstar@gmail.com

Chon, Kyong Hee, Western Kentucky University, Bowling Green, khchon@gmail.com

Chow, Sy-Miin, Penn State University, 413 BBH Building, Penn State University, University Park, Pennsylvania, 16802, symiin@psu.edu

Chu, Man-Wai, CRAME/University of Alberta, Edmonton, Canada

Chung, Meng-ta C., Teachers College, Columbia University, 525 W 120th St, New York, New York, 10027, mc3128@columbia.edu

Circi, Ruhan, University of Colorado Boulder, 1300 30th St. B4/27, Boulder, Colorado, 80303, ruhan.circi@colorado.edu

Clark, Amy K., University of Kansas, 1122 West Campus Road, Rm. 730, Lawrence, 66045, akclark@ku.edu

Clark, J. Michael, Pearson, Tulsa, Oklahoma, mike.clark@pearson.com

Clarke, Marguerite, World Bank, 1818 H St., NW, Washington, District of Columbia, 20433, mclarke2@worldbank.org

Clauser, Jerome, American Board of Internal Medicine, 184 Maiden St, Philadelphia, Pennsylvania, 19127-1406, jclauser@abim.org

Cohen, Yoav, NITE, Jerusalem, Israel

Colvin, Kimberly F., Massachusetts Institute of Technology, 990 Massachusetts Ave., Apt. 50, Arlington, Massachusetts, 02476, colvin@mit.edu

Cook, Gary, Wisconsin Center for Education Research, Madison, Wisconsin

Cook, Howard G., University of Wisconsin, 730 Cledel St, Oregon, Wisconsin, 53575

Cook, Linda, Educational Testing Service, Princeton, New Jersey

Coppola, Nancy, NJIT, Newark, nancy.w.coppola@njit.edu

Croft, Michelle C., ACT, Inc., Iowa City, Iowa, michelle.croft@act.org

Contact Information for Individual and Coordinated Sessions First Authors

Cubbellotti, Stephen, Fordham University, 175 Pinehurst Ave, Apt 5C, New York, New York, 10033, cubbellotti@fordham.edu

Cui, Weiwei, NISS, Research Triangle, wcui@niss.org

Culbertson, Michael J., University of Illinois, Urbana, Illinois, culbert1@illinois.edu

Culpepper, Steve, University of Illinois at Urbana-Champaign, Flossmoor, Illinois, sculpepp@illinois.edu

Cunningham, Paula, University of Iowa, 1133 Summer Street, Grinnell, Iowa, 50112, paula-cunningham@uiowa.edu

Dadey, Nathan, University of Colorado Boulder, Boulder, Colorado, nathan.dadey@colorado.edu

Davidson, Anne, CTB/McGraw Hill, Carson City, U.S.A.

Davis, Laurie L., Pearson, 400 Center Ridge Drive, Austin, Texas, 78753, laurie.davis@pearson.com

Davison, Mark L., University of Minnesota, Education Sciences Building Ste. 250, 56 E. River Road, Minneapolis, Minnesota, 55455, mldavison_2000@yahoo.com

Dawson, Anthony D., Cambridge Assessment, Cambridge, United Kingdom, dawson.a@cie.org.uk

De Champlain, Andre F., Medical Council of Canada, 23 Devon Road, Newtown, Pennsylvania, 18940, adechamplain@mcc.ca

Deane, Paul D., Educational Testing Service, Rosedale Road, MS 11-R, Princeton, New Jersey, 08541, pdeane@ets.org

Debeer, Dries J.L., University of Leuven, Tiensestraat 102, bus 3713, Leuven, 3000, Belgium, dries.debeer@ppw.kuleuven.be

DeBoeur, George, Project 2061, AAAS, Washington, District of Columbia

DeMars, Christine E., James Madison University, MSC 6806, Center for Assessment & Research, Harrisonburg, Virginia, 22807-6806, demarsce@jmu.edu

Deng, Sien, University of Wisconsin-Madison, Madison, Wisconsin, 53705, sdeng7@wisc.edu

Deng, Weiling, ETS, 660 Rosedale Road, Princeton, New Jersey, 08540, wdeng@ets.org

Diakow, Ronli, New York University, New York, rd110@nyu.edu

Diao, Qi, CTB/McGraw-Hill, Monterey, qi_diao@ctb.com

Diaz-Billelo, Elena, Center for Assessment, Dover, New Hampshire

DiBello, Louis V., Learning Sciences Research Institute at University of Illinois at Chicago, Bloomingdale, Illinois, 60108, ldibello@uic.edu

DiCerbo, Kristen, Pearson, Avondale, Arizona, kdicerbo@cox.net

Ding, Cody, University of Missouri-St. Louis, St. Louis, dinghc@umsl.edu

Ding, Shuliang, Jiangxi Normal University, Nanchang, China, ding06026@163.com

Dogan, Enis, PARCC, Washington DC, District of Columbia, edogan@parcconline.org

Doorey, Nancy, K-12 Center at ETS, 4601 Beechwold Avenue, Wilmington, Delaware, 19803

Dorans, Neil, Educational Testing Service, Princeton

Draney, Karen, UC Berkeley, 6633 Armour Dr., Oakland, California, 94611, kdraney@berkeley.edu

Drasgow, Fritz, University of Illinois, Urbana-Champaign, Urbana-Champaign

Du, Yi, Data Recognition Corporation, 5793 Turtle Lake Rd, Shoreview, 551264774, yi.du.lin0020@gmail.com

Dwyer, Andrew C., Castle Worldwide, Inc., Morrisville, North Carolina, acdwyer777@yahoo.com

Edejer, Eilene, Loyola University, Chicago, Illinois

Elliott, Stephen N., Arizona State University, Tempe, Arizona, 85259, steve_elliott@asu.edu

Contact Information for Individual and Coordinated Sessions First Authors

Embretson, Susan, Georgia Institute of Technology, Atlanta, Georgia

Evran, Derya, Rutgers University, Rutgers University GSE, 10 Seminary Place, New Brunswick, New Jersey, 08901, derya.evran@rutgers.edu

Faulkner-Bond, Molly M., University of Massachusetts Amherst, 11 Arnold Avenue, Apt 2B, Northampton, Massachusetts, 01060, mfaulkne@educ.umass.edu

Fay, Derek M., Arizona State University, Tempe, Arizona, derek.fay@asu.edu

Feinberg, Richard A., National Board of Medical Examiners, Philadelphia, rfeinberg@nbme.org

Feldman, Anna, Montclair State University, Montclair, feldmana@mail.montclair.edu

Feng, Gary, Educational Testing Service, 660 Rosedale Rd, Princeton, New Jersey, 08540, gfeng@ets.org

Ferrara, Steve, Pearson, 1242 C St., SE, Washington, District of Columbia, 20003, sferrara1951@gmail.com

Fife, James H., Educational Testing Service, Mail Stop 18-E, 660 Rosedale Road, Princeton, New Jersey, 08541, jfife@ets.org

Fina, Anthony D., The University of Iowa, Iowa City, Iowa, anthony-fina@uiowa.edu

Finch, Holmes, Ball State University, Department of Educational Psychology, Muncie, Indiana, 47306, whfinch@bsu.edu

Fitzpatrick, Joseph, University of Kansas, Lawrence, Kansas, jfitz@ku.edu

Flor, Michael, Educational Testing Service, MS 13-R, Rosedale Road, Princeton, New Jersey, 08541, mflor@ets.org

Flowers, Claudia P., UNC Charlotte, Charlotte, North Carolina, 28205, cpflower@uncc.edu

Foley, Brett P., Alpine Testing Solutions, 10345 SW 119th St., Denton, Nebraska, 68339, brett.foley@alpinetesting.com

Forgione, Pascal D., K-12 Center at ETS, 701 Brazos Street, Suite 500, Austin, Texas, 78701, pdforgione@k12center.org

Forte, Ellen, edCount, Washington, District of Columbia, eforte@edcount.com

Fraillon, Julian, Australian Council for Educational Research, Melbourne, Australia, fraillon@acer.edu.au

French, Brian, Washington State University, Pullman, Washington, 99164, frenchb@wsu.edu

Fu, Yin, University of South Carolina, West Columbia, fuy@email.sc.edu

Fujimoto, Ken A., University of Illinois at Chicago, Chicago, Illinois

Gaertner, Matthew, Pearson, 4105 Love Bird Lane, Austin, Texas, 78730, matthew.gaertner@pearson.com

Gaj, Shameem N., ETS, 660 Rosedale Road, Princeton, New Jersey, 08546, sgaj@ets.org

Galpern, Andrew, University of California, Berkeley, Berkeley, California, 94720, galpern@berkeley.edu

Gándara, M. Fernanda, University of Massachusetts Amherst, Amherst, mgandara@educ.umass.edu

Gao, Lingyun, ACT, Iowa City, lingyun.gao@act.org

Gardner, Emily, World Bank, 1818 H St, NW, Washington, District of Columbia, 20433

Geranpayeh, Ardeshir, University of Cambridge, 1 Hills Road, Cambridge, CB23 7QN, United Kingdom, geranpayeh.a@cambridgeenglish.org

Gierl, Mark, University of Alberta, 6-110 Education North, Faculty of Education, Edmonton, Alberta, T6G2G5, Canada, mark.gierl@ualberta.ca

Giesy, Philip, Renaissance Learning, Inc, Vancouver, phil.giesy@renlearn.com

Glas, Cees, University of Twente, Enschede, Netherlands, lrutkows@indiana.edu

Contact Information for Individual and Coordinated Sessions First Authors

Gonulates, Emre, Michigan State University, 2092 Lac Du Mont Apt D2, Haslett, Michigan, 48840, gonulat1@msu.edu

Gonzalez, Jorge, Department of Statistics, Pontificia Universidad Catolica de Chile, Faculty of Mathematics, Av. Vicuna Mackenna 4860, Macul, Santiago, Chile

Goodman, Joshua, Pacific Metrics, 12517 S. 18th Circle, Jenks, Oklahoma, 74037, jgoodman@pacificmetrics.com

Gorin, Joanna, ETS, 660 Rosedale Road, Princeton, New Jersey, 08540

Graf, Edith A., ETS, Princeton, New Jersey, 08541, agraf@ets.org

Gregory, Kelvin D., Australian Curriculum, Assessment and Reporting Authority, Sydney, Australia, kdg2505@gmail.com

Grochowalski, Joe, Fordham University, New York

Gu, Lixiong, Educational Testing Service, 660 Rosedale Rd., MS 13P, Princeton, New Jersey, 08540, lgu@ets.org

Guerere, Claudia, University of South Florida, 780 7th Ave NE, Largo, Florida, 33770, cguerere@mail.usf.edu

Guo, Fanmin, Graduate Management Admission Council, 11921 Freedon Drive, Suite 300, Reston, Maryland, 20190, fguo@gmac.com

Guo, Hongwen, ETS, Princeton, New Jersey, hguo@ets.org

Guo, Rui, University of Illinois at Urbana Champaign, 610 Oakland Ave Apt 206, Urbana, Illinois, 61802, ruiguo1@illinois.edu

Haertel, Geneva, SRI International, Menlo Park, California, geneva.haertel@sri.com

Hall, Erika, Center for Assessment, Iowa City, Iowa, ehall@nciea.org

Halpin, Peter F., New York University, Steinhardt School of Culture, Education, and Human Development - 82 Washington Square East, New York, New York, 10003, peter.halpin@nyu.edu

Ham, Eun Hye, Michigan State University, 620 Farm Lane (Erickson) CEPSE, East Lansing, 48823, thanks02@gmail.com

Han, Kyung (Chris) T., Graduate Management Admission Council, Reston, Virginia, 20124, khan@gmac.com

Hansen, Eric G., Educational Testing Service, Mailstop 10-R, 660 Rosedale Road, Princeton, New Jersey, 08541, ehansen@ets.org

Hansen, Mark, UCLA, Los Angeles, California, 90025, markhansen@ucla.edu

Hao, Jiangang, Educational Testing Service, Rosedale Rd, Princeton, New Jersey, 08541, jhao@ets.org

Haring, Samuel H., University of Texas, Cedar Park, Texas, 78613, samuel.haring@utexas.edu

Harrell, Lauren, UCLA, Los Angeles, California, laurenharrell@ucla.edu

Harrison, George M., University of Hawaii at Manoa, Honolulu, 96822, georgeha@hawaii.edu

Hauser, Carl H., Northwest Evaluation Association, 10065 SW Flicka Pl, Beaverton, Oregon, 97008, carl.hauser@nwea.org

Hazen, Tim, Iowa Testing Programs, Iowa City, Iowa, timothy-hazen@uiowa.edu

He, Qiwei, Educational Testing Service, 660 Rosedale Road E213, Princeton, New Jersey, 08541, qhe@ets.org

He, Wei, NWEA, Portland, wei.he@nwea.org

He, Yong, ACT, Inc., Iowa City, yong.he@act.org

Heilman, Michael, Educational Testing Service, Princeton, New Jersey, 08541

Henson, Robert A., The University of North Carolina at Greensboro, 305 Stapleton Way, High Point, North Carolina, 27265-1159, rahenson@uncg.edu

Herman, Joan, UCLA/CRESST

Higgins, Derrick, Educational Testing Service, 325 W. School House Lane, Philadelphia, Pennsylvania, 19144, dhiggins@ets.org

Contact Information for Individual and Coordinated Sessions First Authors

Higgins, Jennifer, Measured Progress, Newton

Hillier, Tracey, University of Alberta, Edmonton, Canada

Ho, Andrew, Harvard University, Cambridge, Massachusetts, andrew_ho@gse.harvard.edu

Ho, Emily, Fordham University, 411 East Fordham Road, Dealy Hall 236, Bronx, New York, 10458, eho2@fordham.edu

Hoffman, Erin, GlassLab, Richmond, erin@instituteofplay.org

Holtzman, Kathleen Z., NBME, Philadelphia, Pennsylvania, 19104, kholtzman@nbme.org

Hortensius, Lisabet M., University of Minnesota, 3029 France Avenue S, Apt 212, Minneapolis, Minnesota, 55416, horte005@umn.edu

Hou, Jianlin, The School District of Palm Beach, Wellington,jianlin.hou@palmbeachschools.org

Hou, Likun, American Institute of CPAs, Ewing, New Jersey, lhou@aicpa.org

Hsu, Ying-Ju, University of Iowa, 1001 Oakcrest St Apt24E, Iowa City, Iowa, 52246, ying-ju-hsu@uiowa.edu

Hu, Huiqin, DRC, Plymouth, Minnesota, huiqin_hu@yahoo.com

Hu, Jinxiang, University of Florida, Gainesville, jinxianghu@ufl.edu

Hu, Xueying, Texas A&M University-College Station, 904 University Oaks Blvd, Apt 97, College Station, Texas, 77840, catherine23@neo.tamu.edu

Huang, Lan, University of Minnesota, Minneapolis, Minnesota, huang371@umn.edu

Hutchison, Christine, University of Virginia, 206 Blumont Drive, Lynchburg, Virginia, 24503, cch5k@virginia.edu

Hwang, Hojun, University of Washington, Seattle, Washington, junhwang@uw.edu

Iaconangelo, Charles, Rutgers, the State University of New Jersey, New Brunswick, New Jersey, 08901, charles.iaconangelo@gmail.com

Ilich, Maria O., University of Washington, 514-A S. Catalina Ave., Redondo Beach, California, 90277, moi1@u.washington.edu

Im, Suk Keun, University of Kansas, Lawrence, Kansas, sukkeun@ku.edu

Isac, Maria Magdalena, CRELL, Centre for Research on Lifelong Learning, Via E. Fermi 2749, Ispra, TP 361, Italy, maria-magdalena.isac@jrc.ec.europa.eu

Iseli, Markus R., CRESST UCLA, Los Angeles, iseli@cse.ucla.edu

Jang, Hyesuk, Michigan State University, 2520 Wood St Apt 11, Lansing, Michigan, 48912, janghyes@msu.edu

Jang, Yoonsun, University of Georgia, Athens, Georgia, 30602

Janssen, Rianne, KU Leuven, Dekenstraat 2 (PB 3773), Leuven, 3000, Belgium, rianne.janssen@ppw.kuleuven.be

Jiang, Yanlin, ETS, Princeton, yjiang@ets.org

Jin, Kuan-Yu, Hong Kong Institute of Education, Hong Kong, Hong Kong, kyjin@ied.edu.hk

Joe, Jilliam N., ETS, 660 Rosedale Rd., MSC 18-E, Princeton, New Jersey, 08541, jjoe@ets.org

Johnson, Matthew S., Teachers College, Columbia University, 525 W. 120th Street, Box 118, New York, New York, 10803, johnson@tc.edu

Jurich, Daniel P., James Madison University, Harrisonburg, Virginia, 22801, jurichdp@dukes.jmu.edu

Kang, Hyeon-Ah, University of Illinois at Urbana-Champaign, Room 210. TA Box. 1310 S. 6th St., Champaign, Illinois, 61820

Kaniskan, Burcu, NCBE, Madison, Wisconsin, burcukaniskan@gmail.com

Kannan, Priya, Educational Testing Service, 660 Rosedale Road, Princeton, 08844, pkannan@ets.org

Kao, Shu-chuan, Pearson, 1 N Dearborn St. Suite 1050, Chicago, Illinois, 60602, shu-chuan.kao@pearson.com

Contact Information for Individual and Coordinated Sessions First Authors

Kaplan, David, University of Wisconsin - Madison, Department of Educational Psychology, Madison, dkaplan@education.wisc.edu

Kapoor, Shalini, University of Iowa, Iowa City, Iowa, 52246, shalini-kapoor@uiowa.edu

Karkee, Thakur B., Measurement Inc, Durham, tkarkee@measinc.com

Kelcey, Ben, University of Cincinnati, Cincinnati

Keller, Lisa A, University of Massachusetts Amherst, Amherst, Massachusetts

Keng, Leslie, Pearson, Austin

Kern, Justin L., University of Illinois at Urbana-Champaign, Champaign, Illinois, 61820, kern4@illinois.edu

Ketterlin-Geller, Leanne R., Southern Methodist University, Dallas, lkgeller@smu.edu

Khademi, Abdolvahab, University of Massachusettes Amherst, Amherst, Massachusetts, vahab.khademi@gmail.com

Khoo, Siek Toon, Australian Council for Educational Research, 19 Prospect Hill Rd, Camberwell, Victoria, 3124, Australia, khoo@acer.edu.au

Khorramdel, Lale, Educational Testing Service, Research and Development, 660 Rosedale Road, Princeton, New Jersey, 08541

Kieftenbeld, Vincent, CTB/McGraw-Hill Education, Monterey, vincent.kieftenbeld@ctb.com

Kim, Do-Hong, University of North Carolina at Charlotte, Charlotte, dkim15@uncc.edu

Kim, Jinho, University of California, Berkeley, 2344 Virginia Street, Berkeley, 94709, potatopaul@gmail.com

Kim, Se-Kang, Fordham University, Dept. of Psychology, 441 E. Fordham Rd., Bronx, New York, 10458, sekim@fordham.edu

Kim, Sooyeon, Educational Testing Service, 660 Rosedale Road, MS 08-P, Princeton, New Jersey, 08541, skim@ets.org

Kim, Yoon Jeon, Florida State University, 1114 W. Call Street, 3212, TALLAHASSEE, Florida, 32306, yjkim.fsu@gmail.com

Kim, YoungKoung, The College Board, 45 Columbus Ave, New York, New York, 10023, ykim@collegeboard.org

King, David R., Georgia Institute of Technology, Atlanta

Kingsbury, G. Gage, Independent Consultant to NWEA, 6134 SE Lincoln, Portland, Oregon, 97215, gagekingsbury@comcast.net

Kingston, Neal M., University of Kansas, 1122 West Campus Drive, Room 720a, Lawrence, Kansas, 66047, nkingsto@ku.edu

Kolen, Michael, University of Iowa, Iowa City

Kopp, Jason P., James Madison University, 821 S. Main St., MSC 6806, Harrisonburg, Virginia, 22807

Kuhfeld, Megan, UCLA, 8259 Fountain Ave, Apt 6, West Hollywood, California, 90046, megan.kuhfeld@gmail.com

Lai, Hollis, University of Alberta, Edmonton, Alberta, Canada, hollis.lai@ualberta.ca

Lakin, Joni, Auburn University

LaMar, Michelle M., UC Berkeley, Graduate School of Education, 3659 Tolman Hall, Berkeley, California, 94720, mlamar@berkeley.edu

Lan, Ming-Chih, University of Washington, 312 Miller Box 353600, Seattle, Washington, 98195, mclan@uw.edu

Lane, Suzanne, University of Pittsburgh, 605 Victory Road, Pittsburgh, Pennsylvania, 15237, sl@pitt.edu

Lash, Andrea A., WestEd, 730 Harrison St, San Francisco, California, 94107, alash@wested.org

Lathrop, Quinn N., University of Notre Dame, South Bend, Indiana, qlathrop@nd.edu

Latifi, Syed Muhammad Fahad, University of Alberta, Edmonton, Alberta, T6G 2G5, Canada, fahad.latifi@ualberta.ca

Latour, Thibaud, CRP Henri Tudor, Luxembourg, Luxembourg

Contact Information for Individual and Coordinated Sessions First Authors

Lazendic, Goran, The Australian Curriculum, Assessment and Reporting Authority, 255 Pitt Street, Sydney, New South Wales, 2000, Australia, goran.lazendic@acara.edu.au

Leacock, Claudia, CTB McGraw-Hill, 100 Bleecker Street, Apt 27A, New York, New York, 10012, claudia_leacock@ctb.com

Lee, Guemin, Yonsei University, 50 Yonsei-Ro, Seodaemoon-Ku, Seoul, 120-749, Korea, Republic of, guemin@yonsei.ac.kr

Lee, Han K., University of South Carolina, 1001 True St., Apt. 818, Columbia, 29209, leehk@email.sc.edu

Lee, HyeSun, University of Nebraska-Lincoln, 25 Teachers College Hall, Lincoln, Nebraska, 68588, hyesun.kj.lee@gmail.com

Lee, Jieun, University of Minnesota, Minneapolis, leex3828@umn.edu

Lee, Minji K., UMass Amherst, Minneapolis, minjikang@gmail.com

Lee, Moonsoo, UCLA, Los Angeles, California, mslee91@gmail.com

Lee, Philseok, University of South Florida, Tampa, philseok@mail.usf.edu

Lee, Soo Youn, Rutgers University, New Brunswick, New Jersey, sooyolee@scarletmail.rutgers.edu

Lee, Young-Sun, Teachers College, Columbia University, New York

Lei, Pui-Wa, The Pennsylvania State University, University Park, Pennsylvania, puiwa@psu.edu

Leighton, Jacqueline P., CRAME/University of Alberta, Edmonton, Alberta, T6G 2G5, Canada, jacqueline.leighton@ualberta.ca

Lewis, Charles, Educational Testing Services, Princeton, New Jersey

Lewis, Daniel, CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey, 93940, daniel_lewis@ctb.com

Li, Deping, ETS, Princeton, New Jersey, dli@ets.org

Li, Dongmei, ACT Inc., Iowa City, dongmei.li@act.org

Li, Feifei, Educational Testing Service, Princeton, New Jersey, 08541, fli@ets.org

Li, Isaac, University of South Florida, Tampa, Florida, liy1@mail.usf.edu

Li, Ming, University of Maryland, 9664 Scotch Haven Drive, Vienna, Virginia, 22181, liming@umd.edu

Li, Shuhong, Educational Testing Service, Princeton, sli@ets.org

Li, Tianli, ACT Inc., Iowa City, Iowa, tianli.li@act.org

Li, Tongyun, University of Maryland-College Park, College Park, Maryland, 20742, tongyun@umd.edu

Li, Xiaomin, The Hong Kong Institute of Education, Hong Kong, Hong Kong, nickylxm@yahoo.com.hk

Li, Xin, ACT, Inc., Iowa City, xin.li@act.org

Li, Zhushan, Boston College, Campion Hall, Room 336D, 140 Commonwealth Avenue, Chestnut Hill, 02467, zhushan.li@gmail.com

Lietz, Petra, Australian Council for Educational Research, Lv 3, 97 Pirie Street, Adelaide, South Australia, 5000, Australia, lietz@acer.edu.au

Lim, EunYoung, KICE, Seoul, Korea, Republic of, eylim.ed@gmail.com

Lin, Chih-Kai, University of Illinois at Urbana-Champaign, Champaign, Illinois, cjlin4@illinois.edu

Lin, Haiyan, ACT, Inc., Iowa City, Iowa, haiyan.lin@act.org

Liu, Chen-Wei, The Hong Kong Institute of Education, Hong Kong, Hong Kong, genwei007@gmail.com

Liu, Fu, University of North Carolina at Greensboro, Greensboro, North Carolina, 27403, floraliu11@hotmail.com

Liu, Jinghua, Secondary School Admission Test Board, 862 County Road 518, Skillman, New Jersey, 08543, jliu@ssat.org

Liu, Lei, ETS, Princeton, New Jersey, lliu001@ets.org

Contact Information for Individual and Coordinated Sessions First Authors

Liu, Xiangdong, University Of Iowa, Iowa City, xiangdong-liu@uiowa.edu

Liu, Xin Lucy, Data Recognition Corporation, Maple Grove, Minnesota, lliu@datarecognitioncorp.com

Liu, Yuming, ETS, Pennington, New Jersey, 08534, yliu@ets.org

Lockeman, Kelly S., Virginia Commonwealth University, School of Medicine, PO Box 980466, Richmond, Virginia, 23298-0466, kslockeman@vcu.edu

Lockwood, J.R., Educational Testing Service, 109 Eileen Drive, Pittsburgh, Pennsylvania, 15227, jrlockwood@ets.org

Longabach, Tanya, University Of Kansas, Lawrence, Kansas, 66045, tlongabach@ku.edu

Lord-Bessen, Jennifer, Fordham University, 43-09 40 Street, 4H, Sunnyside, New York, 11104, jlord2@fordham.edu

Lorie, William, Questar Assessment, Inc., Washington, District of Columbia, william.lorie@gmail.com

Loughran, Jessica, University of Kansas, 1122 West Campus Road, CETE, Joseph R. Pearson Hall, Lawrence, Kansas, 66044, jtl@ku.edu

Lu, Ying, Educational Testing Service, Princeton, New Jersey, 08541, ylu@ets.org

Lu, Zhenqiu (Laura), University of Georgia, Athens, Georgia, 30602

Luo, Xin, Michigan State University, East Lansing, Michigan, 48823, luoxin1@msu.edu

Lutz, Megan E., Georgia Institute of Technology, 2459 Hosea L. Williams Dr, Atlanta, Georgia, 30317, megan.lutz@gmail.com

Lyrén, Per-Erik, Umeå University, Umeå, Sweden, per-erik.lyren@edusci.umu.se

Madison, Matthew J., University of Georgia, University of Georgia, Aderhold Hall 125-A, Athens, Georgia, 30602, matthewm@uga.edu

Magis, David, Université de Liège, Liege, Belgium

Mao, Liyang, Michigan State University, 6180 N Hagadorn Rd, Apt 5, East Lansing, Michigan, 48823, maoliyan@msu.edu

Mao, Xia, Pearson, 2339 Cameron Way, Iowa City, Iowa, 52246, xia.mao@pearson.com

Marion, Scott F., Center for Assessment, 31 Mount Vernon St, Dover, New Hampshire, 03820, smarion@nciea.org

Maris, Gunter, CITO - University of Amsterdam, Arnhem, Netherlands, gunter.maris@cito.nl

Masters, James S., Pearson, 8760 Renfrew Street, Powell, 43065, james.masters@pearson.com

Matlock, Ki L., University of Arkansas, Fayetteville, Fayetteville, Arkansas, kilynn@uark.edu

Matos-Elefonte, Haifa, College Board, New York, hmatoselefonte@collegeboard.org

Mattar, John, AICPA, Ewing, New Jersey, 08628

Matthews-Lopez, Joy, National Association of Boards of Pharmacy, 1400 Maplewood Drive, Athens, Ohio, 45701

McCaffrey, Daniel, Educational Testing Service, Pittsburgh, Pennsylvania

McCall, Martha, Smarter Balanced Assessment Consortium, Portland, Oregon, 97219, mccall.marty@gmail.com

McCoy, Thomas P., UNC Greensboro, Greensboro, tpmccoy@uncg.edu

McDonald, Stefanie R., American Institutes for Research, Washington, District of Columbia, stefanie.mcdonald@gmail.com

McGinn, Dan, Harvard University, Cambridge

McGrane, Joshua, University of Western Australia, Perth, Western Australia, Australia

McGuire, Leah, Measured Progress, Dover, mcguire.leah@measuredprogress.org

McJunkin, Linette M., Computerized Assessments and Learning, 2715 Maverick Ln, Lawrence, Kansas, 66046, lmcjunkin@ku.edu

McNeish, Dan, University of Maryland, College Park, 7305 Yale Ave, College Park, Maryland, 20740, dmcneish@umd.edu

Contact Information for Individual and Coordinated Sessions First Authors

McTavish, Thomas S., Pearson, Denver, Colorado, tom.mctavish@pearson.com

Mee, Janet, National Board of Medical Examiners, 3750 Market Street, Philadelphia, 19104, jmee@nbme.org

Melendez, Carlos R., Educational Testing Service (ETS), Princeton, New Jersey, cmelendez@ets.org

Mendelovits, Juliette, Australian Council for Educational Research, Private Bag 55, Camberwell, Victoria, 3124, Australia, juliette.mendelovits@acer.edu.au

Meng, Huijuan, Pearson, 7719 Tanglewood Court, Edina, Minnesota, 55439, huijuan.meng@pearson.com

Merriman, Jennifer, The College Board, Newtown, Pennsylvania, 18940, jmerriman@collegeboard.org

Meyer, Patrick, University of Virginia, P.O. Box 400265, 405 Emmet Street South, Charlottesville, Virginia, 22901, meyerjp@virginia.edu

Meyers, Jason, Pearson, Austin

Miel, Shayne, Measurement Incorporated, 2202 Chapel Hill Rd., Durham, North Carolina, 27707, smiel@measinc.com

Mikolajetz, Anna, University of Jena, Jena, Germany, anna.mikolajetz@uni-jena.de

Miller, Tamara B., University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, 53217, millertb@uwm.edu

Morrison, Kristin, Georgia Institute of Technology, Atlanta, Georgia, kmorrison3@gatech.edu

Moustaki, Irini, London School of Economics

Murphy, Daniel L., Pearson, Austin, dan.murphy@pearson.com

Natesan, Prathiba, University of North Texas, Denton, Texas, 76201, prathiba.natesan@unt.edu

Naumann, Alexander, German Institute for Internationel Educational Research (DIPF), Solmsstraße 73, Frankfurt am Main, 60486, Germany, naumanna@dipf.de

Nenkova, Ani, University of Pennsylvania, Philadelphia, Pennsylvania, 19104, nenkova@seas.upenn.edu

Newton, Paul E., Institute of Education, University of London, 20 Bedford Way, London, WC1H 0AL, United Kingdom, newton.paul@virginmedia.com

Nichols, Paul, Pearson, Iowa City, Iowa, paul.nichols@pearson.com

Nydick, Steven, Pearson VUE, Minneapolis, Minnesota, nydic001@umn.edu

Olivera Aguilar, Margarita, Educational Testing Service, Princeton, New Jersey, 08648, molivera-aguilar@ets.org

Olsen, James B., Rensissance Learning, 1781 Sunrise Drive, Orem, Utah, 84097, jamesbolsen@hotmail.com

Oranje, Andreas, ETS, Princeton, aoranje@ets.org

Oswald, Marc, Open Assessment Technologies S.A., Esch-sur-Alzette, 4362, Luxembourg, marc@taotesting.com

Öztürk Gübes, Nese, Hacettepe University, Beytepe Kampüsü, Çankaya, Ankara, 06800, Turkey, neseozturk07@gmail.com

Paek, Youngshil, Univ. of Illinois at Urbana-Champaign, Champaign, yspaek@hotmail.com

Pak, Seohong, University of Iowa, Coralville, Iowa, seohong-pak@uiowa.edu

Park, Jiyoon, Federation of State Boards of Physical Therapy, Alexandria, jpark@fsbpt.org

Park, Yoon Soo, University of Illinois at Chicago, Chicago, yspark2@uic.edu

Patton, Jeffrey M., University of Notre Dame, Notre Dame, Indiana, jpatton1@nd.edu

Perlin, Rachel, CUNY Graduate Center, New York, New York, rperlin@gc.cuny.edu

Contact Information for Individual and Coordinated Sessions First Authors

Peterson, Jaime, University of Iowa, Iowa City, Iowa, jaime-peterson@uiowa.edu

Pivovarova, Margarita, University of Toronto, Department of Economics, 150 St. George Street, Toronto, Ontario, M5S3G7, Canada, rita.pivovarova@mail.utoronto.ca

Porter, Andrew, University of Pennsylvania, University of Pennsylvania Graduate School of Education, 3700 Walnut Street, Philadelphia, Pennsylvania, 19104

Powers, Sonya, Pearson, Austin, sopowers@gmail.com

Puhan, Gautam, ETS, Princeton, New Jersey, 08541, gpuhan@ets.org

Qian, Hong, National Council of State Boards of Nursing, 1000 Main Street, Apt.1A, Evanston, Illinois, 60202, qianhon1@gmail.com

Qian, Jiahe, Educational Testing Service, MS 02-T, Princeton, New Jersey, 08541, jqian@ets.org

Qian, Xiaoyu, ETS, Princeton, xqian@ets.org

Qiu, Xue-Lan, Hong Kong Institute Of Education, Hong Kong, Hong Kong, xlqiu@ied.edu.hk

Quellmalz, Edys, WestEd, Redwood City, California

Ramineni, Chaitanya, ETS, Rosedale Road, MS 18-E, Princeton, New Jersey, 08541, cramineni@ets.org

Raymond, Mark, NBME, Philadelphia, Pennsylvania, 19104, mraymond@nbme.org

Reckase, Mark, Michigan State University, East Lansing, Michigan, 48824

Reddy, Linda, Rutgers University, New Brunswick

Reshetar, Rosemary A., The College Board, 661 Penn Street, Suite B, Newtown, Pennsylvania, 18940, rreshetar@collegeboard.org

Rho, Yun Jin, Pearson, Boston, yjdemian@gmail.com

Rios, Joseph A., University of Massachusetts Amherst, 103 South Street, Apt. 302, Northampton, Massachusetts, 01060, jarios@educ.umass.edu

Roduta Roberts, Mary, University of Alberta, 2-64 Corbett Hall, Department of Occupational Therapy, Faculty of Rehabilitation Medicine, University of Alberta, Edmonton, Alberta, T6G 2G4, Canada, mroberts@ualberta.ca

Rogat, Aaron, ETS, Princeton, New Jersey

Rogers, Jane, University of Connecticut, Dept of Educational Psychology, 249 Glenbrook Road, U-3064, Storrs, Connecticut, 06269, jane.rogers@uconn.edu

Rojas, Guaner, Universidad de Costa Rica, San José, Costa Rica, guaner.rojas@ucr.ac.cr

Roohr, Katrina C., Educational Testing Service, Princeton, New Jersey, kroohr@ets.org

Rotou, Ourania, Educational Testing Service, Princeton, New Jersey

Ruby, Alan, University of Pennsylvania, 3451 Walnut St, Philadelphia, Pennsylvania, 19104

Rutkowski, Leslie, Indiana University, 201 N. Rose Ave, Bloomington, Indiana, 47405, lrutkows@indiana.edu

Rutstein, Daisy, SRI International, 2349 Menzel Place, Santa Clara, California, 95050, daisy.rutstein@sri.com

Saito, Hidetoshi, Ibaraki University, Mito, Japan, cldwtr@mx.ibaraki.ac.jp

Samonte, Kelli M., University of North Carolina Greensboro, 1721 Remington Point Court, Walkertown, North Carolina, 27051, kelli.samonte@gmail.com

Sato, Edynn, Pearson, San Francisco, California, edynn.sato@pearson.com

Scalise, Kathleen, University of Oregon, 5267 University of Oregon, Eugene, Oregon, 97403, kscalise@uoregon.edu

Schofield, Lynne S., Swarthmore College, 500 College Avenue, Swarthmore, 19081, lschofi1@swarthmore.edu

Schulz, Wolfram, Australian Council for Educational Research, 19 Prospect Hill Road, Camberwell, Victoria, 3124, Australia, wolfram.schulz@acer.edu.au

Contact Information for Individual and Coordinated Sessions First Authors

Schwarz, Richard, ETS, Salinas, rschwarz@ets.org

Scott, Sarah M., University of California, Merced, Merced, California, 95348, sscott7@ucmerced.edu

Semmelroth, Carrie, Boise State University, Boise, Idaho, 83725-1725, carriesemmelroth@boisestate.edu

Seo, Daeryong, Pearson, San Antonio, Texas, 78232, daeryong.seo@pearson.com

Seo, Dong Gi, Michigan Department of Education, Lansing, Michigan, 48864, seod@michigan.gov

Seol, Jaehoon, American Institute of CPAs, Ewing

Sha, Shuying, University of North Carolina at Greensboro, 201 Revere Dr, Apt9, Greensboro, North Carolina, 27407, s_sha@uncg.edu

Shang, Yi, John Carroll University, University Heights, Ohio, yshang@jcu.edu

Sheehan, Kathleen M., ETS, Princeton, New Jersey, ksheehan@ets.org

Shermis, Mark D., The University of Akron, 213 Crouse Hall, Akron, Ohio, 44718, mshermis@uakron.edu

Shin, Ahyoung, University Of Iowa, Iowa City, ahyoung-shin@uiowa.edu

Shin, David, Pearson, Iowa City

Shin, Hyo Jeong, University of California, Berkeley, Berkeley, hjshin@berkeley.edu

Shin, MinJeong, University of Massachusetts Amherst, Amherst, minjeong@educ.umass.edu

Shu, Lianghua, CTB/McGraw-Hill, 4904 Sea Crest Ct, Seaside, California, 93955, lianghua_shu@ctb.com

Sie, Haskell, Pennsylvania State University, 3161 Sheller's Bend, State College, Pennsylvania, 16803, mathzlovers@yahoo.com

Sinharay, Sandip, CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey, California, 93940, sandip_sinharay@ctb.com

Sireci, Stephen, University of Massachusetts Amherst, 156 Hills House South, School of Education, Amherst, Massachusetts, 01003, sireci@acad.umass.edu

Skorupski, William P., University of Kansas, 1122 West Campus Rd., 639 JRP, Lawrence, Kansas, 66045, wps@ku.edu

Socha, Alan, James Madison University, 799 Quince Orchard Blvd., Apartment 12, Gaithersburg, Maryland, 20878, sochaab@dukes.jmu.edu

Song, Tian, Pearson, 5604 E Galbraith Rd, Cincinnati, Ohio, 45236, tian.song@pearson.com

Stahl, John A., Pearson, 1 North Dearborn St, Chicago, Illinois, 60602, john.stahl@pearson.com

Steiner, Peter, University of Wisconsin, Madison, Wisconsin

Stenner, Alfred J., MetaMetrics, Inc., 1000 Park Forty Plaza Drive, Suite 120, Durham, North Carolina, 27713, jstenner@lexile.com

Stephenson, Shonte, GlassLab, Redwood City

Stevens, Shawn, University of Michigan, Ann Arbor

Stevenson, Claire E., Leiden University, Psychology Methods & Statistics, Postbus 9555, Leiden, 2300 RB, Netherlands, cstevenson@fsw.leidenuniv.nl

Stoddart, Laura, DRC, Maple Grove, Minnesota, LStoddart@datarecognitioncorp.com

Stone, Elizabeth, Educational Testing Service, Princeton, estone@ets.org

Stopek, Joshua, AICPA, Ewing, New Jersey, 08628

Stuart, Elizabeth, Johns Hopkins, Baltimore, Maryland

Su, Ihui, University of Missouri, 16 Hill Hall, Columbia, Missouri, 65211

Su, Ya-Hui, National Chung Cheng University, No.168, Sec. 1, University Rd., Min-Hsiung Township, Chia-yi County, 621, Taiwan, psyyhs@ccu.edu.tw

Contact Information for Individual and Coordinated Sessions First Authors

Su, Yu-Lan, ACT, Iowa City, Iowa, 52241, suyulan@gmail.com

Sun, Guo Wei, National Sun Yat-Sen University, No. 70, Lienhai Rd., Kaohsiung, 80424, Taiwan, hiroko1210@gmail.com

Sussman, Joshua M., University of California Berkeley, 4511 Tolman Hall, Berkeley, 94720, jsussman@berkeley.edu

Swaminathan, Hariharan, University of Connecticut, Neag School of Education 3064, 249 Glenbrook Road, Storrs, Connecticut, 06269, swami@uconn.edu

Swanson, David B., NBME, Philadelphia, Pennsylvania, 19104, dswanson@nbme.org

Swerdzewski, Peter, Regents Research Fund, New York City

Swygert, Kimberly, National Board of Medical Examiners, 3750 Market Street, Philadelphia, Pennsylvania, 19104, kswygert@nbme.org

Tang, Shuwen, University of Wisconsin-Milwaukee, 3941 N Downer Ave, Milwaukee, Wisconsin, 53211, tangsw.1106@gmail.com

Tang, Wei, University of Alberta, 4220, 139 Ave, Edmonton, T5Y 2Y2, Canada, wtang3@ualberta.ca

Tannenbaum, Richard J., ETS, Princeton, rtannenbaum@ets.org

Tao, Shuqin, Data Recognition Corporation, 15134 83rd Place N, Maple Grove, Minnesota, 55311, shuqin.tao@gmail.com

Thissen, David, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, dthissen@email.unc.edu

Thissen-Roe, Anne, Comira, San Mateo, athissenroe@comiratesting.com

Thompson, Tony, ACT, Inc., Iowa City, Iowa, tony.thompson@act.org

Thompson, Vanessa M., Georgia Institute of Technology

Tian, Wei, Naional Assessment Of Educational Quality, Beijing, 100875, China, tianwei65396@163.com

Tijmstra, Jesper, Utrecht University, Utrecht, Netherlands, j.tijmstra@gmail.com

Timms, Michael J., Australian Council for Educational Research, ACER, 19 Prospect Hill Road, Camberwell, Victoria, 3124, Australia, mike.timms@acer.edu.au

Tindal, Gerald, University of Oregon, BRT-175 Education, 5262 University of Oregon, Eugene, Oregon, 97403, geraldt@uoregon.edu

Topczewski, Anna M., Pearson, 3505 Edgewood Drive, Ann Arbor, Michigan, 48104, topczewski.anna@gmail.com

Torres Irribarra, David, UC Berkeley, 744 Euclid Avenue, Berkeley, 94708, david@torresirribarra.me

Traynor, Anne, Michigan State University, 55 Crescent Rd, Apt 1524L, East Lansing, Michigan, 48823, traynor2@msu.edu

Trierweiler, Tammy J., Prometric, 1200 Lenox Drive, Suite 100, Lawrenceville, New Jersey, 08648, tjtrier@gmail.com

Tudor, Joshua, The University of Iowa, Iowa City, Iowa, joshua-tudor@uiowa.edu

Turner, Ronna C., University of Arkansas, 250 Graduate Education Building, Fayetteville, Arkansas, 72701, rcturner@uark.edu

Valdivia Vazquez, Juan A., Washington State University, Pullman, Washington, 99163

van den Heuvel, Jill, Alpine Testing Solutions, Hatfield, jill.vandenheuvel@alpinetesting.com

van der Linden, Wim J., CTB/McGraw-Hill, 20 Ryan Ranch, Monterey, California, 93940, wim_vanderlinden@ctb.com

van Rijn, Peter, ETS Global, Amsterdam, Netherlands, pvanrijn@etsglobal.org

VanIwaarden, Adam, University of Colorado, Boulder, Colorado

Vermeulen, Jorine, University of Twente, Ringbaan West 170-16, Tilburg, 5041LT, Netherlands, jorine.vermeulen@cito.nl

Contact Information for Individual and Coordinated Sessions First Authors

von Davier, Alina A., Educational Testing Service, Rosedale Rd. MS 02 T, Princeton, New Jersey, 08541, avondavier@ets.org

von Davier, Matthias, ETS, Rosedale Road, Princeton, New Jersey, 08541, mvondavier@ets.org

Waldman, Marcus, Teachers College, Columbia University, New York, mrw2152@tc.columbia.edu

Walker, Michael, ETS, Princeton, New Jersey

Wan, Lei, Pearson, Coraville, lei.wan@pearson.com

Wang, Changjiang, Pearson, Iowa City, changjiang.wang@pearson.com

Wang, Chun, University of Minnesota, Minneapolis, Minnesota, wang4066@umn.edu

Wang, Keyin, Michigan State University, Room 410 Erickson Hall, East Lansing, Michigan, 48823, keyinw0323@gmail.com

Wang, Min, the University of Iowa, Iowa City, min-wang@uiowa.edu

Wang, Shiyu, University of Illinois at Urbana-Champaign, Urbana, Illinois, wsybianbian@gmail.com

Wang, Shudong, NWEA, 121 NW Everett Street, Portland, Oregon, 97029, shudong.wang@nwea.org

Wang, Wenhao, University of Kansas, Lawrence, wwh8623@gmail.com

Wang, Wenyi, Jiangxi Normal University, 99 Ziyang Dadao, Nanchang, 330022, China, wenyiwang2009@gmail.com

Wang, Xi, University of Massachusetts Amherst, 279 Amherst Rd, Apt. 52, Sunderland, Massachusetts, 01375, xiw@educ.umass.edu

Wang, Xiaojing, University of Connecticut, Department of Statistics, University of Connecticut, 215 Glenbrook RD, U-4120, Storrs, Connecticut, 06269

Wang, Xiaolin, Indiana University Bloomington, Bloomington, Indiana, coolxlw@gmail.com

Wang, Xinrui, Pearson VUE, 3603 N Kedvale Ave, Unit 3, Chicago, Illinois, 60641, xinrui.wang@pearson.com

Warner, Zachary B., New York State Education Department, Albany, New York

Way, Denny, Pearson, Iowa City, Iowa

Weeks, Jonathan, Educational Testing Service, Rosedale Road, MS E13, Princeton, New Jersey, 08541, jweeks@ets.org

Wei, Hua, Pearson, 4535 Jewel Ln. N., Plymouth, Minnesota, 55446, hua.wei@pearson.com

Wei, Youhua, ETS, Princeton, ywei@ets.org

Wells, Craig, University of Massachusetts Amherst, Amherst, cswells@educ.umass.edu

Welsh, Megan, University of Connecticut, U-3064, 249 Glenbrook Rd., Storrs, Connecticut, 06249, megan.welsh@uconn.edu

Wen, Yao, University of Wisconsin Milwaukee, Milwaukee, Wisconsin, 53211, yaowen@uwm.edu

Westrick, Paul A., ACT, 500 ACT Drive, P.O. Box 168, Iowa City, Iowa, 52243-0168, paul.westrick@act.org

White, John, SAS Institute, Cary

Whithaus, Carl, UC, Davis, Davis, cwwhithaus@ucdavis.edu

Wihardini, Diah, UC Berkeley, 3659 Tolman Hall, Berkeley, California, 94720-1670, diah.wihardini@berkeley.edu

Williams, Frank E., Educational Testing Service, Princeton, New Jersey, fwilliams001@ets.org

Williams, Immanuel, Rutgers, The State University of New Jersey, New Brunswick, New Jersey

Williams, Natasha, Pearson, Austin, Texas

Williamson, David, Educational Testing Service, Princeton, New Jersey, dmwilliamson@ets.org

Williamson, Gary L., MetaMetrics, Inc., 1000 Park Forty Plaza Drive, Suite 120, Durham, North Carolina, 27713, gwilliamson@lexile.com

Contact Information for Individual and Coordinated Sessions First Authors

Wilson, Mark R., University of California, Berkeley, 227 Trinity Avenue, Kensington, California, 94708, markw@berkeley.edu

Wolf, Raffaela, University of Pittsburgh, Pittsburgh, raw59@pitt.edu

Wolfe, Edward W., Pearson, Iowa City, Iowa, ed.wolfe@pearson.com

Wolkowitz, Amanda A., Alpine Testing Solutions, Chesterfield, amanda.wolkowitz@alpinetesting.com

Wollack, James A., University of Wisconsin-Madison, Madison

Wong, Cheow Cher, Singapore Examinations and Assessment Board, 298 Jalan Bukit Ho Swee, Singapore, Singapore, wong_cheow_cher@seab.gov.sg

Wright, Dan B., ACT Inc., Austin, Texas, daniel.wright@act.org

Wu, Haiyan, National Board of Osteopathic Medical Examiners, 1001 E 3rd At Apt 1, Chicago, Illinois, 55805, hw07d@my.fsu.edu

Wu, Hao, Boston College, McGuinn Hall 300, 140 Commonwealth Ave, Chestnut Hill, Massachusetts, 02467, schpnhr@gmail.com

Wu, Pei-Chen, University of South Florida, Tampa

Wyse, Adam E., Michigan Department of Education, Arden Hills, Minnesota, wysea@michigan.gov

Xie, Chao, University of Maryland-College Park, 9308 Cherry Hill Rd Apt 816, College Park, Maryland, 20740, olisha.chao@gmail.com

Xiong, Yao, Penn State U, State College, Pennsylvania, yzx110@psu.edu

Xu, Jing-Ru, Michigan State University, East Lansing, xujingru@msu.edu

Xu, Ting, University of Pittsburgh, Drexel Hill, tix3@pitt.edu

Xu, Yuning, Arizona State University, Tempe, Arizona, yuningxu@asu.edu

Yang, Fan, The University of Iowa, 206 6Th St Apt C3, Coralville, Iowa, 52241, fan-yang-3@uiowa.edu

Yang, Ji Seung, University of Maryland, College Park, yangjsedu@gmail.edu

Yang, Ping, University of Missouri, Columbia, Missouri, pyq3b@mail.missouri.edu

Yang, Zhiming, Educational Testing Service, 158 Hartford Lane, Newtown, Pennsylvania, 18940, yzm506jx@yahoo.com

Yao, Lihua, DMDC, 400 Gigling RD, Seaside, California, 93955, lihua.yao.civ@mail.mil

Ye, Meng, Institute of Developmental Psychology, Beijing Normal University, China., Beijing, China

Ye, Sam, UIUC, Champaign

Yel, Nedim, Arizona State University, 615 S Hardy Dr Apt 36, Tempe, Arizona, nedim@asu.edu

Yilmaz, Mustafa, University of Kansas Center for Educational Testing and Evaluation, Lawrence, Kansas, myilmaz@ku.edu

Yoon, Jiyoung, Seoul Women's Univ., Seoul, Korea, Republic of, ellie5900@naver.com

Yoon, Su-Youn, Educational Testing Service, 660 Rosedale Road, Princeton, New Jersey, 08541, syoon@ets.org

Yu, Lei, Questar Assessment Inc., Apply Valley, lyu@questarai.com

Zahner, Doris, CAE, New York, dzahner@cae.org

Zapata-Rivera, Diego, Educational Testing Service, 660 Rosedale Road MS-16R, Princeton, New Jersey, 08541, dzapata@ets.org

Zara, Anthony, Pearson VUE, 5601 Green Valley Drive, Bloomington, Minnesota, 55437, tony.zara@pearson.com

Zenisky, April L., University of Massachusetts Amherst, 119 Lathrop Street, South Hadley, Massachusetts, 01075, azenisky@educ.umass.edu

Zhang, Changhui, ACT, Inc., Iowa City, Iowa

Zhang, Jiahui, Michigan State University, 1450 Spartan Vlg Apt E, East Lansing, Michigan, 48823, zhang321@msu.edu

Contact Information for Individual and Coordinated Sessions First Authors

Zhang, Jinming, University of Illinois at Urbana-Champaign, Champaign, Illinois, jmzhang@illinois.edu

Zhang, Liru, Delaware Department of Education, 401 Federal Street, Suite 2, Dover, Delaware, 19901, liru.zhang@doe.k12.de.us

Zhang, Litong, CTB, Monterey, litong_zhang@ctb.com

Zhang, Mengyao, University of Iowa, Iowa City, mengyao-zhang@uiowa.edu

Zhang, Mingcai, Michigan State University, 4382 Okemos Road G102, Okemos, Michigan, 48864, zhangmc@msu.edu

Zhang, Mo, Educational Testing Service, Princeton, New Jersey, mzhang@ets.org

Zhang, Ting, American Institutes for Research, 438 Ridge Rd. Apt. 11, Greenbelt, Maryland, 20770, rainyzt@yahoo.com

Zhao, Fei, Arizona Department of Education, Phoenix, Arizona, fei.zhao@azed.gov

Zheng, Chanjin, UIUC, Champaign, russelzheng@gmail.com

Zheng, Chunmei, Pearson-Always Learning, 895 Maplewood Drive, Coralville, Iowa, zhengchunmei5@gmail.com

Zheng, Qiwen, University of Maryland, College Park

Zheng, Yi, University of Illinois at Urbana-Champaign, 1310 S. Sixth Street, Education Bldg, Rm 210, Champaign, Illinois, 61820, yizheng1@illinois.edu

Zhou, Xuechun, Pearson, San Antonio, zhouxuec@msu.edu

Zhou, Yan, Indiana University Bloomington, Bloomington, zhou25@indiana.edu

Zhu, Mengxiao, ETS, Princeton, New Jersey, mzhu@ets.org

Zimmermann, Judith, ETH Zurich, Zurich, Switzerland, judith.zimmermann@inf.ethz.ch

Zu, Jiyun, Educational Testing Service, Princeton, jzu@ets.org

Zwick, Rebecca, Educational Testing Service, Santa Barbara, California, 93108, rzwick@cox.net

NCME 2014 • Schedule-At-A-Glance

Time	Room	Type	ID	Title
Wednesday, April 2, 2014				
8:00 a.m.-5:00 p.m.	Commonwealth B	TS	AA	flexMIRT®: Flexible Multilevel Multidimensional Item Analysis and Test Scoring
8:00 a.m.-5:00 p.m.	Commonwealth D	TS	BB	Assessing Soft Skills: K-12 to Higher Education, Domestic and International
8:00 a.m.-5:00 p.m.	Washington A	TS	CC	Testing Accommodations for Computer-Administered, Postsecondary Readiness Assessments
8:00 a.m.-5:00 p.m.	Commonwealth A	TS	DD	Diagnostic Measurement: Theory, Methods and Applications
8:00 a.m.-5:00 p.m.	Washington B	TS	EE	Application of Principled Design and Development in Large-Scale Assessment
8:00 a.m.-12:00 noon	Commonwealth C	TS	FF	A Graphical and Nonlinear Mixed Model Approach to IRT
8:00 a.m.-12:00 noon	Washington C	TS	GG	Introduction to Natural Language Processing in Educational Measurement
1:00 p.m.-5:00 p.m.	Commonwealth C	TS	HH	A Practitioner's Guide to Growth Models
1:00 p.m.-5:00 p.m.	Washington C	TS	II	Using Visual Displays to Inform Assessment Development and Validation
4:00 p.m.-7:00 p.m.	Jefferson Boardroom			Board of Directors Meeting
Thursday, April 3, 2014				
8:00 a.m.-5:00 p.m.	Commonwealth A	TS	JJ	Analyzing NAEP Data Using Direct Estimation Approach with AM
8:00 a.m.-5:00 p.m.	Commonwealth B	TS	KK	Multidimensional Item Response Theory: Theory and Applications and Software
8:00 a.m.-5:00 p.m.	Commonwealth C	TS	LL	Test Equating Methods and Practices
8:00 a.m.-5:00 p.m.	Commonwealth D	TS	MM	Bayesian Networks in Educational Assessment
8:00 a.m.-12:00 noon	Washington B	TS	NN	Combining Fun, Learning, and Measurement: Game Development in the Classroom and Beyond
8:00 a.m.-12:00 noon	Washington C	TS	OO	Effective Item Writing for Valid Measurement
8:30 a.m.-11:30 a.m.	Washington A	TS	PP	Writing Effective Research Reports

AWR=Award Winning Research • CS=Coordinated Session
EB-CS=Electronic Board Coordinated Session • EB-PS=Electronic Board Paper Session
IS=Invited Session • PS=Paper Session • TS=Training Session
**Unless otherwise noted, sessions will take place at the Loews Hotel.*

1:00 p.m.-5:00 p.m.	Washington A	TS	QQ	Landing Your Dream Job for Graduate Students
1:00 p.m.-5:00 p.m.	Washington B	TS	RR	Software Development Meets Statistics/ Measurement Research
1:00 p.m.-5:00 p.m.	Washington C	TS	SS	TAO Workshop: A Hands-on Introduction to TAO, the Open Source Assessment Platform
Friday, April 4, 2014				
8:00 a.m.-9:40 a.m.	Washington	CS	A1	Applications of Bayesian Networks in Education
8:00 a.m.-9:40 a.m.	Regency A	PS	A2	Constructing Adaptive Tests and Assessing Fit Characteristics
8:00 a.m.-9:40 a.m.	Regency B	PS	A3	Diagnostic Models: Polytomous Data and Attributes
8:00 a.m.-9:40 a.m.	Regency C1	PS	A4	Angoff Judgments, Procedures, Scores and Exercises
8:00 a.m.-9:40 a.m.	Regency C2	PS	A5	Testlets, Dependence, Carry-Over
8:00 a.m.-9:40 a.m.	Commonwealth A	PS	A6	Standard Errors
8:00 a.m.-9:40 a.m.	Commonwealth B	CS	A7	Measuring "Hard-to-Measure" Aspects of Text Complexity
8:00 a.m.-9:40 a.m.	Commonwealth C	PS	A8	Aberrant Item Responses
8:00 a.m.-9:40 a.m.	Commonwealth D	PS	A9	Multidimensionality Issues
9:40 a.m.-10:00 a.m.	Regency Ballroom Foyer			Refreshment Break
10:00 a.m.-11:40 a.m.	Washington	IS	B1	21st Century Skills Debate
10:00 a.m.-11:40 a.m.	Regency A	IS	B2	A Look at Our Psychometric History: Contributions of International Scholars
10:00 a.m.-11:40 a.m.	Regency B	CS	B3	Scoring Related Challenges and Solutions in Technology-Enabled Assessments
10:00 a.m.-11:40 a.m.	Regency C1	PS	B4	Topics in Measurement Gaps: The Need to Know
10:00 a.m.-11:40 a.m.	Regency C2	PS	B5	Model Fit Statistics / Propensity Score Matching
10:00 a.m.-11:40 a.m.	Commonwealth A	IS	B6	Joint NATD and NCME Symposium
10:00 a.m.-11:40 a.m.	Commonwealth B	PS	B7	Measuring Digital Learning Across Countries and Over Time
10:00 a.m.-11:40 a.m.	Commonwealth C	PS	B8	Reliability Issues
10:00 a.m.-11:40 a.m.	Commonwealth D	PS	B9	Multidimensional Models for Polytomous Data

11:50 a.m.-12:40 p.m.	Washington	IS	C1	Measuring Learner Engagement with Data Mining
11:50 a.m.-12:40 p.m.	Regency B	IS	C2	True Grit
12:50 p.m.-2:20 p.m.	Millennium	EB-PS	1	Bayesian Networks
12:50 p.m.-2:20 p.m.	Millennium	EB-PS	2	Classification Models and Approaches
12:50 p.m.-2:20 p.m.	Millennium	EB-PS	3	DIF Identification, Causes, and Consequences
12:50 p.m.-2:20 p.m.	Millennium	EB-PS	4	DIF: Specifics on Power, Purification, Guessing, and Criteria
12:50 p.m.-2:20 p.m.	Millennium	EB-PS	5	Aspects of Student Growth Estimation
12:50 p.m.-2:20 p.m.	Millennium	EB-PS	6	Score Reports: What is New on the Horizon?
12:50 p.m.-2:20 p.m.	Millennium	EB-CS	7	Subscore Reporting in Adaptive Testing
12:50 p.m.-2:20 p.m.	Millennium	EB-CS	8	Validity Arguments for Measures of Teacher Effectiveness
12:50 p.m.-2:20 p.m.	Millennium	EB-PS	9	Value-Added Models, Methods and Analysis
12:50 p.m.-2:20 p.m.	Millennium	EB-CS	10	Experimental Research on Test Item Design and Construction
2:30 p.m.-4:10 p.m.	Washington	CS	D1	Meeting the Challenges to Measurement in an Era of Accountability
2:30 p.m.-4:10 p.m.	Regency A	PS	D2	Scoring and Exposure in Adaptive Settings
2:30 p.m.-4:10 p.m.	Regency B	CS	D3	Three Ways to Improve Subscores Using Multidimensional Item Response Theory”
2:30 p.m.-4:10 p.m.	Regency C1	PS	D4	Solutions in the Development of Technology-Enhanced Items
2:30 p.m.-4:10 p.m.	Regency C2	PS	D5	Compromised Items and Invalid Data
2:30 p.m.-4:10 p.m.	Commonwealth A	PS	D6	Difficulties in Realizing Equivalence and Equating
2:30 p.m.-4:10 p.m.	Commonwealth B	CS	D7	Integrating Games, Learning, and Assessment
2:30 p.m.-4:10 p.m.	Commonwealth C	PS	D8	Anchoring and Equating
2:30 p.m.-4:10 p.m.	Commonwealth D	PS	D9	Longitudinal Models
4:05 p.m.-6:05 p.m.	PCC, Level 200, Room 204A	AERA		Standards for Educational and Psychological Testing: Major Changes and Implications to Users
4:20 p.m.-6:00 p.m.	Washington	PS	E1	Featured Contributions
4:20 p.m.-6:00 p.m.	Regency A	CS	E2	Improving Technical Quality of Computerized Adaptive Testing in K-12 Assessments

AWR=Award Winning Research • CS=Coordinated Session
EB-CS=Electronic Board Coordinated Session • EB-PS=Electronic Board Paper Session
IS=Invited Session • PS=Paper Session • TS=Training Session
**Unless otherwise noted, sessions will take place at the Loews Hotel.*

4:20 p.m.-6:00 p.m.	Regency B	PS	E3	Some Big Ideas in Innovative Measures: Ubiquitous Assessment, National Norms, Learning Progressions, Networked Reading Comprehension
4:20 p.m.-6:00 p.m.	Regency C1	PS	E4	Mobile Devices: Results of Assessment on Tablets
4:20 p.m.-6:00 p.m.	Regency C2	CS	E5	How Comparable are International Comparisons of Educational Outcomes?
4:20 p.m.-6:00 p.m.	Commonwealth A	CS	E6	Modern Profile Analysis via Multivariate Statistics (PAMS)
4:20 p.m.-6:00 p.m.	Commonwealth B	CS	E7	Extensions to Evidence Based Standard Setting
4:20 p.m.-6:00 p.m.	Commonwealth C	PS	E8	Multilevel Models
4:20 p.m.-6:00 p.m.	Commonwealth D	PS	E9	Solutions for Difficult Problems and Unusual Data
6:30 p.m.-8:00 p.m.	Howe Room			NCME and AERA Division D Reception
Saturday, April 5, 2014				
7:45 a.m.-9:00 a.m.	Regency B			NCME Breakfast and Business Meeting
9:00 a.m.-9:40 a.m.	Regency B			Presidential Address
10:00 a.m.-11:40 a.m.	Commonwealth A	IS	F1	NCME Book Series: New NCME Applications of Educational Measurement and Assessment Book Series
10:00 a.m.-11:40 a.m.	Commonwealth D	AWR	F2	Award Winning Research
10:00 a.m.-11:40 a.m.	Regency C1	CS	F3	Evaluating the Evaluations: Exploring Improvements to Measuring Classroom Mathematics Instruction
10:00 a.m.-11:40 a.m.	Regency C2	PS	F4	Item and Rater Drift
10:00 a.m.-11:40 a.m.	Washington A	PS	F5	Goodness of Fit Statistics / Propensity Score Matching
10:00 a.m.-11:40 a.m.	Commonwealth B	CS	F6	Designing System of Next Generation Science Assessments: Challenges, Choices, Trade-Offs
10:00 a.m.-11:40 a.m.	Commonwealth C	CS	F7	What is the Best Way to Use the Term 'Validity'?
10:00 a.m.-11:40 a.m.	Washington B	PS	F8	Missing Data
11:50 a.m.-12:40 p.m.	Washington	IS	G1	A Location Scale Item Response Theory (IRT) Model for Ordinal Questionnaire Data
11:50 a.m.-12:40 p.m.	Regency B	IS	G2	Model-Based Tools Embedded Within Games to Assess and Support Important Competencies

AWR=Award Winning Research • CS=Coordinated Session
EB-CS=Electronic Board Coordinated Session • EB-PS=Electronic Board Paper Session
IS=Invited Session • PS=Paper Session • TS=Training Session
**Unless otherwise noted, sessions will take place at the Loews Hotel.*

12:50 p.m.-2:20 p.m.	Millennium Hall	EB-PS	11	Method and Model Violations and Distortions
12:50 p.m.-2:20 p.m.	Millennium Hall	EB-PS	12	Item Parameter Approaches
12:50 p.m.-2:20 p.m.	Millennium Hall	EB-PS	13	Software Comparisons
12:50 p.m.-2:20 p.m.	Millennium Hall	EB-PS	14	Assessing and Measuring All Students
12:50 p.m.-2:20 p.m.	Millennium Hall	EB-PS	15	Automated Scoring in Multiple Contexts
12:50 p.m.-2:20 p.m.	Millennium Hall	EB-PS	16	Collaboration and Assessment: Co-Constructing Understanding
12:50 p.m.-2:20 p.m.	Millennium Hall	EB-CS	17	Decision Support and Measurement Insight Through Interactive Statistical Visualization
12:50 p.m.-2:20 p.m.	Millennium Hall	EB-CS	18	Advances in Data Forensic Methodologies
12:50 p.m.-2:20 p.m.	Millennium Hall	EB-CS	19	Cognitive Diagnostic Models to Address Issues in Next Generation Assessments
2:30 p.m.-4:10 p.m.	Washington A	PS	H1	Multiple Attempts Assessment or Innovative Score Reporting
2:30 p.m.-4:10 p.m.	Washington B	PS	H2	Comparison Studies
2:30 p.m.-4:10 p.m.	Regency A	PS	H3	Grand Bargain of CAT: Lessons Learned, Field Testing, Challenges & Pool Design
2:30 p.m.-4:10 p.m.	Regency B	CS	H4	Investigating Genre for Writing Measurement and Automated Writing Evaluation
2:30 p.m.-4:10 p.m.	Regency C1	CS	H5	Using Process Data to Assess Student Writing
2:30 p.m.-4:10 p.m.	Regency C2	PS	H6	Pre-Equating and Equivalent Group Equating
2:30 p.m.-4:10 p.m.	Commonwealth A	CS	H7	Causal Modeling When Covariates are Measured With Error
2:30 p.m.-4:10 p.m.	Commonwealth B	PS	H8	Inferences About Student and Teacher Preparation Derived from Measures of Student Learning
2:30 p.m.-4:10 p.m.	Commonwealth C	PS	H9	Bifactor and Multidimensional Models
2:30 p.m.-4:10 p.m.	Commonwealth D	PS	H10	Q Matrix Issues
2:30 p.m.-4:10 p.m.	Millennium Hall		H11	Graduate Student Poster Session
4:20 p.m.-6:00 p.m.	Washington A	CS	I1	Equal Intervals and Vertical Scales: Figments of a Psychometric Imagination?
4:20 p.m.-6:00 p.m.	Washington B	IS	I2	An Introduction to Bayesian Modeling with Applications to Educational Measurements
4:20 p.m.-6:00 p.m.	Regency A	PS	I3	Adaptive Item Selection
4:20 p.m.-6:00 p.m.	Regency B	PS	I4	Measuring Teacher Effectiveness and Educator Performance

AWR=Award Winning Research • CS=Coordinated Session
EB-CS=Electronic Board Coordinated Session • EB-PS=Electronic Board Paper Session
IS=Invited Session • PS=Paper Session • TS=Training Session
**Unless otherwise noted, sessions will take place at the Loews Hotel.*

4:20 p.m.-6:00 p.m.	Regency C1	CS	I5	Spiraling Contextual Questionnaires in Educational Large-Scale Assessments
4:20 p.m.-6:00 p.m.	Regency C2	PS	I6	Equating Methods and Consequences
4:20 p.m.-6:00 p.m.	Commonwealth A	CS	I7	Exploring Some Solutions for Measurement Challenges in Automated Scoring
4:20 p.m.-6:00 p.m.	Commonwealth B	CS	I8	Human Scoring Behavior and the Interplay Between Humans and Machines
4:20 p.m.-6:00 p.m.	Commonwealth C	PS	I9	Cognitive Diagnostic Model Fit
4:20 p.m.-6:00 p.m.	Commonwealth D	PS	I10	Human Scoring Issues
Sunday, April 6, 2014				
5:45 p.m.-7:00 p.m.	Loews Lobby			NCME Fun Run and Walk
8:00 a.m.-9:40 a.m.	Washington	CS	J1	Scoring Issues for Next-Generation Performance Assessments: An Example from CBAL
8:00 a.m.-9:40 a.m.	Regency A	PS	J2	Multidimensional and Two-Tier Models in Adaptive and Multistage Tests
8:00 a.m.-9:40 a.m.	Regency B	IS	J3	Diversity and Testing Issues
8:00 a.m.-9:40 a.m.	Regency C1	PS	J4	Through-Course and Interim Assessment: Growth and Comparisons with Summative Results
8:00 a.m.-9:40 a.m.	Regency C2	PS	J5	Speededness
8:00 a.m.-9:40 a.m.	Commonwealth A	CS	J6	Longitudinal and Vertical Equating in the Australian National Assessment Program
8:00 a.m.-9:40 a.m.	Commonwealth B	PS	J7	Mathematics Assessment
8:00 a.m.-9:40 a.m.	Commonwealth C	PS	J8	Cheating Detection
8:00 a.m.-9:40 a.m.	Commonwealth D	PS	J9	Item Calibration
9:40 a.m.-10:00 a.m.	Regency Ballroom Foyer			Refreshment Break
10:00 a.m.-11:40 a.m.	Washington	IS	K1	NCME Career Award Lecture and Debate on Cognitive Approaches in Educational Measurement
10:00 a.m.-11:40 a.m.	Regency A	IS	K2	Test Fairness: How Will the Revised AERA/APA/NCME Standards Affect Practice?”
10:00 a.m.-11:40 a.m.	Regency B	CS	K3	Detecting and Minimizing Sources of Construct-Irrelevant Variance in Performance Assessments
10:00 a.m.-11:40 a.m.	Regency C1	CS	K4	Linear Logistic Test Model and Item Variation
10:00 a.m.-11:40 a.m.	Regency C2	PS	K5	Measurement Invariance and DIF

AWR=Award Winning Research • CS=Coordinated Session
EB-CS=Electronic Board Coordinated Session • EB-PS=Electronic Board Paper Session
IS=Invited Session • PS=Paper Session • TS=Training Session
**Unless otherwise noted, sessions will take place at the Loews Hotel.*

10:00 a.m.-11:40 a.m.	Commonwealth A	IS	K6	Dynamic Modeling Approaches
10:00 a.m.-11:40 a.m.	Commonwealth B	CS	K7	Leveraging Multiple Perspectives to Develop Technology-Enhanced, Scenario-Based Assessments
10:00 a.m.-11:40 a.m.	Commonwealth C	PS	K8	Items: Types, Parameters, Numbers
10:00 a.m.-11:40 a.m.	Commonwealth D	PS	K9	Providing and Evaluating 21st-Century Test Accommodations
11:50 a.m.-12:40 p.m.	Washington	IS	L1	Criterion-Referenced Measurement: A Half-Century Wasted?
11:50 a.m.-12:40 p.m.	Regency B	IS	L2	The Economic Value of Teacher Quality
12:50 p.m.-2:20 p.m.	Millennium Hall	EB-PS	20	Equity and Equating
12:50 p.m.-2:20 p.m.	Millennium Hall	EB-PS	21	Scaling and Growth
12:50 p.m.-2:20 p.m.	Millennium Hall	EB-PS	22	Mapping and Alignment
12:50 p.m.-2:20 p.m.	Millennium Hall	EB-PS	23	Growth Percentiles/Validity Issues
12:50 p.m.-2:20 p.m.	Millennium Hall	EB-CS	24	Advances in Open-Source Assessment Technology: Benefits and Applications
12:50 p.m.-2:20 p.m.	Millennium Hall	EB-PS	25	Cheating, Scoring, Scaling and Classifying in Multistage, CAT and Technology-Enhanced Assessment
12:50 p.m.-2:20 p.m.	Millennium Hall	EB-PS	26	Item Selection and Content Distribution in Adaptive Testing
12:50 p.m.-2:20 p.m.	Millennium Hall	EB-CS	27	Reporting Student Growth: Issues and Future Score Reporting Efforts
12:50 p.m.-2:20 p.m.	Millennium Hall	EB-PS	28	Automated Scoring: Sampler of Applications
12:50 p.m.-2:20 p.m.	Millennium Hall	EB-PS	29	Analyzing Data from Collaborations: Structure, Models, Applications
12:50 p.m.-2:20 p.m.	Millennium Hall	EB-CS	30	Revising Assessment Systems Around the World: Evolution, Revolution, or Both?
2:30 p.m.-4:10 p.m.	Washington	CS	M1	Developments in Statistical and Psychometric Modeling for International Large-Scale Assessments
2:30 p.m.-4:10 p.m.	Regency A	PS	M2	Calibration and Balancing Information
2:30 p.m.-4:10 p.m.	Regency B	PS	M3	Time Elements in Adaptive Tests
2:30 p.m.-4:10 p.m.	Regency C1	PS	M4	Diagnostic and Formative Assessments: Modeling Strategies
2:30 p.m.-4:10 p.m.	Regency C2	PS	M5	Single Item Scale/Computer & Paper/ Jackknifed Variance Estimation

AWR=Award Winning Research • CS=Coordinated Session
EB-CS=Electronic Board Coordinated Session • EB-PS=Electronic Board Paper Session
IS=Invited Session • PS=Paper Session • TS=Training Session
**Unless otherwise noted, sessions will take place at the Loews Hotel.*

2:30 p.m.-4:10 p.m.	Commonwealth A	CS	M6	Theories of Action About Performance Assessment: Assessment Consortia Validation Research
2:30 p.m.-4:10 p.m.	Commonwealth B	PS	M7	Cut-Scores, Subsets and Smoothing in Standard Setting
2:30 p.m.-4:10 p.m.	Commonwealth C	PS	M8	DIF Analysis Methods
2:30 p.m.-4:10 p.m.	Commonwealth D	PS	M9	Subscores and Error Information
4:00 p.m.-7:00 p.m.	Jefferson Boardroom			NCME Board of Directors Meeting
4:20 p.m.-6:00 p.m.	Washington	CS	N1	English-Learner Measurement: Score Interpretation, Performance Standards and Redesignation Impact
4:20 p.m.-6:00 p.m.	Regency A	CS	N2	Using Enhanced CAT Designs to Improve Future Implementations
4:20 p.m.-6:00 p.m.	Regency B	CS	N3	Automatic Scoring of Non-Traditional Forms of Assessment
4:20 p.m.-6:00 p.m.	Regency C1	PS	N4	Observation Instruments and Rating
4:20 p.m.-6:00 p.m.	Regency C2	PS	N5	Accuracy and Reliability of Diagnostic Information and Raters
4:20 p.m.-6:00 p.m.	Commonwealth A	CS	N6	Practical Approaches to Defining the Test Construct and Specifications
4:20 p.m.-6:00 p.m.	Commonwealth B	CS	N7	Strengthening Student Assessment in Developing Countries
4:20 p.m.-6:00 p.m.	Commonwealth C	PS	N8	Special DIF Analysis
4:20 p.m.-6:00 p.m.	Commonwealth D	PS	N9	Diagnostic Assessment and Classification