001. Chan Zuckerberg Fellows

8:00 to 5:00 pm

Hilton San Francisco Union Square: Yosemite A

Chan Zuckerberg Fellows

002. Classroom Assessment Task Force

8:00 to 5:00 pm

Hilton San Francisco Union Square: Yosemite B

Classroom Assessment Task Force

003. Exploring, Visualizing, and Modeling Big Data with R

Training Session *8:00 to 5:00 pm*

Hilton San Francisco Union Square: Franciscan Ballroom A

Working with big data requires a particular suite of data analytics tools and advanced techniques, such as machine learning (ML). Many of these tools are readily and freely available in R. This full-day session will provide participants with a hands-on training on how to use data analytics tools and machine learning methods available in R to explore, visualize, and model big data. The first half of the session will focus on organizing (manipulating and summarizing) and visualizing (both statically and dynamically) big data in R. The second half will involve a series of short lectures on ML techniques (decision trees, random forest, and support vector machines), as well as hands-on demonstrations applying these methods in R. Examples will be drawn from various assessments (e.g., PISA and TIMSS). Participants will get opportunities to work through several, directed labs throughout the day. The target audience for this session includes graduate students, researchers interested in analyzing big data from large-scale assessments and surveys, and practitioners working with big data on a daily basis. Some familiarity with the R programming language is required. Participants should bring a laptop with R and RStudio installed to be able to complete the labs during the session.

Session Organizers:

Okan Bulut, University of Alberta

Christopher D. Desjardins, St. Michael's College

004. Assessment of Intrapersonal/Interpersonal Skills for K-12, Higher Education, and the Workplace

Training Session *8:00 to 5:00 pm*

Hilton San Francisco Union Square: Union Square 1 & 2

This full-day workshop provides training, discussion, and hands-on experience in developing methods for assessing, scoring, and reporting on interpersonal and intrapersonal skills, for K-12, higher education, and the workplace. Workshop focuses on (a) reviewing the most important skills based on current research and frameworks; (b) general methods for writing good items; (c) standard and innovative measurement methods, including self- and others'-ratings, forced-choice (rankings), anchoring vignettes, and situational judgment testing (SJT); (d) classical and item-response theory (IRT; e.g., 2PL, partial credit, nominal response model) scoring procedures; (e) reliability from classical test theory and IRT; and (f) reporting. Workshop sessions will be organized around item types (e.g., forced-choice, anchoring vignettes). Examples will be drawn from various assessments (e.g., CORE, PISA, NAEP, SuccessNavigator, FACETS). There will be hands-on demonstrations using R for scoring anchoring vignettes and SJTs, and group participation in a collaborative problem solving task. The workshop is designed for a broad audience of assessment developers and psychometricians, working in applied or research settings. Participants should bring laptops preferably with R and Rstudio installed (but help will be provided if needed, and it will be possible to participate as an observer in a group).

Presenters:

Jiyun Zu, Educational Testing Service

Jiangang Hao, Educational Testing Service

Session Organizer:

Patrick Charles Kyllonen, Educational Testing Service

005. Modeling Writing Process Using Keystroke Logs

Training Session *8:00 to 12:00 pm*

Hilton San Francisco Union Square: Union Square 15 & 16

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

Presenters:

Mo Zhang, ETS
Hongwen Guo, Educational Testing Service
Xiang Liu, Educational Testing Service

Session Organizer:

Mo Zhang, ETS

006. An Intuitive Introduction to Maximum Likelihood and Bayesian Estimation for Applied Researchers.

Training Session *8:00 to 12:00 pm*

Hilton San Francisco Union Square: Union Square 17 & 18

The standard two-course statistical methods sequence typically covers through linear regression, ANOVA, and closely related models. However, consulting a text on other common methods (e.g., logistic regression, loglinear models, factor analysis, and item response theory) goes beyond that as they are typically analyzed using maximum likelihood. Even worse for many practitioners, the use of maximum likelihood in these texts often seems to assume the reader has had two semesters of calculus based statistical theory. Similar roadblocks occur when Bayesian methods are introduced. This course, which requires no previous calculus knowledge, introduces maximum likelihood and Bayesian methods for applied researchers. Using the normal distribution, logistic regression, and ability estimation in item response theory (no experience needed!) as examples, the concepts of maximum likelihood and Bayesian estimation are explored graphically and by using participants' laptops with easy to run interactive programs and simulations. Examples from the literature and the methods' limitations are discussed throughout the course.

Session Organizer:

Brian Habing, University of South Carolina

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

Training Session *8:00 to 12:00 pm*

Hilton San Francisco Union Square: Union Square 19 & 20

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

Session Organizer:

Seung Choi, The University of Texas at Austin

008. Python, Machine Learning, and Applications - A Gentle Introduction

Training Session *8:00 to 12:00 pm*

Hilton San Francisco Union Square: Union Square 22

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

Session Organizer:

Zhongmin Cui, ACT, Inc.

009. Using the R package PLmixed to Estimate Complex Measurement and Growth Models

Training Session *8:00 to 12:00 pm*

Hilton San Francisco Union Square: Union Square 23 & 24

Latent variable modeling, including factor analysis, item response theory (IRT), and multilevel modeling, has proved to be an indispensable tool for the analysis of complex educational and behavioral science data. In this workshop, we detail a general framework for formulating complex measurement and growth models within a general latent variable framework, as well as introduce the R package PLmixed (Jeon & Rockwood, 2017) for fitting such models. After providing an introduction to the PLmixed syntax, as outlined by Rockwood and Jeon (2019), we will formulate and fit a number of example models that allow us to explore interesting educational and behavioral science phenomena. These include a multi-rater measurement model, where many teachers provide ratings of many students; a longitudinal multilevel IRT model with crossed random effects to account for students transitioning from middle to high school; a nonlinear growth model; and an explanatory IRT model. By the end of the session, attendees should feel comfortable transitioning from their research question to an appropriate model, to the correct PLmixed syntax for fitting the model. The intended audience includes faculty, graduate students, and practitioners. Participants are encouraged to download R and the PLmixed package from CRAN prior to the workshop.

Presenter:

Minjeong Jeon, UCLA

Session Organizer:

Nicholas John Rockwood, Loma Linda University

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

Training Session *8:00 to 5:00 pm*

Hilton San Francisco Union Square: Union Square 3 & 4

Cognitive diagnostic models (CDMs) are a popular psychometric framework for providing educators and researchers with a fine-grained assessment of students' skills. The purpose of this training session is to introduce participants to two emerging research areas. First, confirmatory applications of CDMs require expert knowledge in the form of a Q matrix to define the latent structure. Research shows that using inaccurate expert knowledge diminishes the performance of confirmatory applications of CDMs. In order to overcome this challenge, we introduce participants to exploratory methods for inferring the latent structure, validate expert knowledge, and support theory development. We review common measurement models, discuss model identifiability, and review parameter estimation. Second, significant interest centers on the use of CDMs in longitudinal settings to track student learning and to understand the process of skill acquisition. We review the challenges with longitudinal applications of CDMs and introduce participants to several flexible models for describing attribute transitions over time. Our session offers a balance of lecture on new methods and real-data applications. In particular, we disseminate and provide participants access to examples involving recently developed open-source R packages to estimate the latent structure and model learning.

Presenters:

Jeffrey Douglas, University of Illinois at Urbana-Champaign
Shiyu Wang, University of Georgia
Yinghan Chen, University of Nevada Reno
James Balamuta, University of Illinois at Urbana-Champaign
Susu Zhang, Columbia University
Yawei Shen, The University of Georgia
Hulya Yigit, University of Illinois at Urbana-Champaign
Auburn Jimenez, University of Illinois at Urbana-Champaign

Session Organizer:

Steven Culpepper, University of Illinois at Urbana-Champaign

011. An Introduction to the Use of Telemetry Data in Video Game Analyses

Training Session *8:00 to 12:00 pm*

Hilton San Francisco Union Square: Union Square 5 & 6

Participants will be introduced to the analysis of video game data with a focus on deriving meaningful measures from player interaction data. A suite of learning games, developed by PBS KIDS to specifically teach concepts of measurement to preschool children, will be used throughout the training session to provide hands-on play experience and cognitive demands analysis. The game will provide a real-world example for data analyses, and a context for telemetry design and best practices. This introductory session will be of interest to people interested in using games for measurement purposes but who lack experience in the area. The training session will be divided into three parts. Part 1: Extracting Meaningful Events and Measures from Gameplay Data will offer hands-on experience with the critical analytical process involved in the identification of important events and the derivation of measures. Part II: Examples of Measures and Analyses of Gameplay Data will focus on basic data analyses approaches that can be used to make sense of gameplay data. Part III: Best Practices From a Game Developer's Perspective will provide a software development perspective on how to instrument games to capture meaningful events. The games require an iPad; a few iPads will be provided.

Presenters:

Elizabeth J. K. H. Redman, UCLA CRESST
Tianying Feng, UCLA/CRESST
Jeremy D Roberts, PBS KIDS Digital

Session Organizer:

Gregory K. W. K. Chung, UCLA/CRESST

012. LNIRT: joint modeling of responses (accuracy) and response times (speed)

Training Session *1:00 to 5:00 pm*

Hilton San Francisco Union Square: Union Square 15 & 16

Large-scale testing programs in educational measurement often use response accuracy (RA) and response time (RT) data to make inferences about test taker's ability and speed, respectively. Computer-based testing offers the possibility to collect item-response time information, by recording the total time spent on each item. Together with the RA data, this kind of information can be used in test design to make more profound inferences about response behavior of the candidates. When following the popular modeling framework of van der Linden (2007) and Klein Entink, Fox, and van der Linden (2009) joint models are constructed by connecting an IRT model with an RT model thereby defining a relationship between the person and item parameters. These joint models have been successfully applied in educational measurement. To extend the psychometric inferences for response accuracy and response times a marginal joint modeling approach is discussed, in which test taker's latent variables for ability and speed are integrated out. This approach will be very advantageous, when attention is focused on identifying changes in ability, speed, and their relationship, and evaluating statistical hypotheses. Examples are given to illustrate the new modeling framework. Model extensions to include other types of process data are discussed.

Session Organizer:

Jean-Paul Fox, University of Twente

013. Item Response Theory and Scale Linking with jMetrik 5.0

Training Session *1:00 to 5:00 pm*

Hilton San Francisco Union Square: Union Square 17 & 18

jMetrik is an open source program for psychometrics. Version 5.0 introduces new features that facilitate research and operational work. In this interactive training session, participants will learn to use jMetrik for IRT and scale linking. Short introductions to a topic will be followed by hands-on experience running the software. Topics include: JMLE and the Rasch family of models; MMLE and the two-, three-, and four-parameter, item response models, and the generalized partial credit model; person scoring; scale linking with Haebaera, and Stocking-Lord procedures; IRT true and observed score equating; and response probability maps. Version 5.0 runs from CSV files, and features a new and concise command syntax. IRT analysis can be executed with three short lines of code. Participants will receive a free copy of the user manual that details the command syntax. The session is suitable for graduate students and researchers who would like to know more about IRT and how to conduct an analysis. Participants with a rich knowledge of IRT will also benefit as they learn how to use jMetrik to conduct simulation studies and integrate it with other statistical software to streamline operational work. Participants will need a laptop. Software and data files will be provided.

Session Organizer:

Patrick Meyer, NWEA

014. Computerized Multistage Adaptive Testing: Theory and Applications (Book by Chapman and Hall)

Training Session *1:00 to 5:00 pm*

Hilton San Francisco Union Square: Union Square 19 & 20

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

Session Organizers:

Duanli Yan, ETS
Alina von Davier, ACTNext
Kyung Han, GMAC

015. Automated Report Generation: SAS and R Officer

Training Session *1:00 to 5:00 pm*

Hilton San Francisco Union Square: Union Square 22

Teams that perform research or summarize student data are often tasked with generating customized reports that display similar information for multiple states and school districts. Instead of manually copying text and data visualizations into a document, researchers and practitioners may take advantage of methods developed in both SAS and R to automatically generate multiple customized versions of a document in a single process. Automating this process not only minimizes the potential for human error, it also capitalizes efficiencies for reoccurring and batch reports. This session will demonstrate the capability of R and SAS to automatically generate reports, technical documentation, research manuscripts, and presentations. Participants will learn how to create document templates (in Word, Excel and PowerPoint), automate such forms of output as text, tables, and figures, and use various coding techniques to improve efficiency. The workshop will review factors to maximize efficiency and avoid potential roadblocks and pitfalls and will incorporate hands-on coding exercises. Additionally, presenters will provide demonstrations of more advanced techniques and provide sample code for later reference. Participants interested in hands-on exercises should bring their own laptop and download SAS University Edition and/or R software prior to the workshop.

Session Organizers:

Elizabeth Patton, Curriculum Associates
Lucinda Simmons, Curriculum Associates

016. Response Similarity Analysis Using R

Training Session *1:00 to 5:00 pm*

Hilton San Francisco Union Square: Union Square 23 & 24

In an era of high-stakes testing, maintaining the integrity of test scores has become an important issue and another aspect of validity. One aspect of investigating testing irregularities is screening item response data for unusual response similarity among test takers. A number of response similarity indices exist in the literature to identify unusual response similarity; however, the computational tools available for practitioners are very limited. This training workshop will introduce the computational foundations of a number of response similarity indices and an R package to compute them. The training will include a 45-min lecture about the computational details of the response similarity indices, a 90-min live demonstration of response similarity analysis using the provided R scripts and datasets, and a 90-min of hands-on activity. The intended audience is testing professionals interested in screening testing data for unusual response similarity. Participants will need to bring a laptop for their own use. The training will assume the participants have a basic understanding of how R works. At the end of the training, participants will be comfortable in using the CopyDetect package in R to compute various response similarity indices and interpret the output.

Session Organizer:

Cengiz Zopluoglu, University of Miami

017. Automated Test Assembly: Optimization and Heuristic Approaches

Training Session *1:00 to 3:00 pm*

Hilton San Francisco Union Square: Union Square 5 & 6

Automated test assembly applies computer algorithms to select items based on content and statistical constraints, and greatly improves the efficiency in test development. In this workshop, we will teach how to implement two ATA algorithms—weighted deviation model (WDM) and mixed integer programming (MIP)—and apply them to assemble linear tests, multistage (MST) panels, linear on the fly tests (LOFT), and shadow tests and item pools in computerized adaptive testing (CAT). For each method we will include: (a) a lecture on the theories behind the algorithm, (b) a live coding session where instructors demonstrate the implementation and applications of the algorithm; and (c) a coding clinic session where instructors help participants analyze and improve their code to assignments. We will use R for MIP and Microsoft Excel for WDM. The intended audience is psychometric practitioners and researchers who need to assemble large, complex linear or adaptive tests. Participants should have basic R and Excel skills. After attending the workshop, participants will (a) understand the theories related to WDM and MIP, (b) be able to apply WDM and MIP to assemble various types of tests, and (c) understand the different goals of ATA for different test models.

Presenter:

Kirk Alan Becker, Pearson

Session Organizer:

Xiao Luo, Measured Progress

018. Automated Test Assembly and More Fun with Optimization Algorithms in R

Training Session *3:30 to 5:30 pm*

Hilton San Francisco Union Square: Union Square 5 & 6

Mixed-integer programming (MIP) algorithms can be used to solve a host of operational problems in an efficient, optimized, and automated fashion, including automated test assembly (ATA). However, optimization software is often prohibitively expensive or difficult to implement and program. Effectively utilizing open-source optimization software requires a thorough understanding of programming and the mechanics of the software, which can be intimidating for many practitioners. In this workshop, participants will learn how to use the free ompr package within R freeware to solve various optimization problems. Examples will include assigning examiners to locations based on stated preferences, balancing severity across examiner teams, assigning reviewers to papers, a simple ATA example, and assembling oral examination case rosters. Participants will complete the workshop with sufficient introductory knowledge and resources to build optimization algorithms for their organizations or research purposes.

Presenter:

Derek Sauder, James Madison University

Session Organizers:

Jason P Kopp, American Board of Surgery
Paulius Satkus, James Madison University

019. Uni/multi-dimensional Computerized Adaptive Testing (CAT): Installation to Administration

Training Session *8:00 to 5:00 pm*

Hilton San Francisco Union Square: Franciscan Ballroom A

After this session, the participants will be able to: 1. Install a CAT framework to their servers / computers 2. Create an item bank and HTML template 3. Code a uni/multidimensional CAT algorithm with R 4. Run their own CAT application This session will focus on Concerto framework, and mirtCAT & Shiny to create CAT applications. We'll code a real-time/live CAT application which is ready for administration based on unidimensional and multidimensional item response theory (IRT) models. Since it's easier to adopt a code from polytomous to dichotomous; we'll work on polytomous IRT models in this session. All the item samples and parameters will be provided before the meeting. After this session; participants will be able to develop their own CAT application and administer it to the examinees related to the context. This session suits better for the PhD students who study on psychometrics, educational / psychological measurement or other related areas. All participants will need to bring their laptops and install their servers (e.g. Digitalocean or Amazon Web Services [AWS]). Participants should also be familiar with coding on R.

Session Organizers:

MURAT DOĞAN ŞAHİN, no
Eren Can Aybek, no

020. Test Equating: Methods and Practices - NCME Training Session

Training Session *8:00 to 5:00 pm*

Hilton San Francisco Union Square: Union Square 1 & 2

See attached file

Session Organizer:

Michael Kolen, University of Iowa

021. Qualitative Coding of Assessment Attributes: An Introduction to Feature Analysis

Training Session *8:00 to 12:00 pm*

Hilton San Francisco Union Square: Union Square 15 & 16

Participants will be introduced to feature analysis, a process of qualitatively coding tasks (such as assessment items or game tasks) against a set of attributes, and a brief overview of quantitative analyses used to determine how these attributes contribute to task performance. This process ensures assessment validity by going beyond simple task description, and by yielding explanations for possible areas of development, identifying task elements suitable for instruction, and providing a method for comparability across tasks. Participants will have the opportunity to practice coding tasks—both traditional multiple-choice math and English language arts items, as well as educational digital games (that may be used as assessments). Participants will learn about test item features that inform future test development, interpretation of test scores, and potential for item bias. Features of digital educational games that inform the creation of educational games and the interpretation of gameplay behavior will also be presented. This session will be of interest to test developers seeking to broaden their understanding of how test item features beyond content/domain interact with student performance, to practitioners and policymakers interested in making inferences about test scores beyond content/domain, and game developers interested in creating games for learning and measurement purposes.

Session Organizers:

Jenny C. Kao, UCLA CRESST
Elizabeth J. K. H. Redman, UCLA CRESST
Gregory K. W. K. Chung, UCLA/CRESST

022. Tools and Strategies for the Design and Evaluation of Score Reports

Training Session *8:00 to 12:00 pm*

Hilton San Francisco Union Square: Union Square 17 & 18

Score reports are often the primary means by which score users receive information about tests. Different score users can substantially vary in their needs for information and their knowledge of the assessment. It is important that score reports, as primary communication tools, are developed so that the results presented are easy to understand and so that they support appropriate inferences for the intended score user. This training session is intended to offer practitioners the tools, strategies, and best practices they need to design and evaluate score reports that are useful and interpretable by stakeholders in different contexts. In this session, we will use various practical hands-on activities, interspersed with lecture, to help participants engage in the practical aspects of designing score reports. We will also equip the attendees with various hand-outs that include summaries of results from research with various audiences and include some effective tools and strategies for designing audience-centric score reports. Participants should bring their own laptops equipped with Microsoft PowerPoint to engage in the practical hands-on sessions. This session could be supplemented by the recent NCME Book on "Score Reporting Research and Applications".

Presenters:

Diego Zapata, ETS
Priya Kannan, Educational Testing Service
April Zenisky, University of Massachusetts Amherst
Sharon C Slater, Educational Testing Service
Gavin Thomas Lumsden Brown, The University of Auckland

Session Organizer:

Priya Kannan, Educational Testing Service

023. Using School-Level Data from the Stanford Education Data Archive

Training Session *8:00 to 10:00 am*

Hilton San Francisco Union Square: Union Square 19 & 20

The Stanford Education Data Archive is a growing, publicly-available database of academic achievement and educational contexts. The nationally-comparable achievement data is based on roughly 330 million standardized test scores for students in nearly every U.S. public school in third through eighth grade from the 2008-09 through 2015-16 school years. To date, SEDA has included only estimates of school district and county-level achievement. In August 2019, the data will be expanded to include estimates of average school-level achievement and will be accessible to a broader audience through a new, interactive website. This workshop is intended to introduce researchers of all levels, practitioners, and policymakers to the new school-level SEDA data. We will provide an overview of both the contents of the SEDA database and the statistical and psychometric methods used to construct the database, focusing on the new school-level data. The workshop will include presentations by the instructors and hands-on activities designed to help users engage directly with the school-level data. All attendees should bring a laptop in order to engage in the activities. Attendees who are interested in using the data for research purposes should have statistical software (e.g., R or Stata) installed on their computers.

Presenter:

Benjamin R Shear, University of Colorado Boulder

Session Organizers:

Sean F. Reardon, Stanford University
Andrew Ho, Harvard Graduate School of Education
Erin M. Fahle, St. John's University

024. Planning and Conducting Standard Setting

Training Session *8:00 to 12:00 pm*

Hilton San Francisco Union Square: Union Square 22

Standard setting doesn't just happen; it takes months of planning and preparation. The goal of this presession is to help participants plan and prepare with confidence for their next standard setting. Its three modules address what to do before, during, and after a formal standard-setting activity, and focus on personnel, materials, schedules, and logistics, with a special emphasis on working with stakeholders and policymakers. While this presession does not focus on any particular standard-setting procedure, it does attend to special needs associated with some of those procedures, including any online or computer-based procedure. The four modules are based on 25 years of practical experience in standard setting and include real-life examples from many of those activities. Participants will receive sample forms, schedules, and planning guides and will be able to role play various aspects of the planning and conduct of standard setting. Method of instruction will be a combination of lecture (PowerPoint presentation), group discussion, and role play. The intended audience includes those responsible for conducting standard setting as well as those responsible for contracting for and overseeing standard-setting activities. Although participants will be given digital handouts, their use of computing devices during the presentation is not required.

Session Organizer:

Michael Bunch, Measurement Incorporated

025. Test Accessibility: What Psychometricians and Test Developers Need to Know

Training Session *8:00 to 10:00 am*

Hilton San Francisco Union Square: Union Square 23 & 24

This training session provides psychometricians and test developers background knowledge and practical information necessary to guide decisions at all stages of the test development cycle. Starting with the Standards for Educational and Psychological Testing (2014) and legal imperatives, session leaders will illustrate how accessibility influences all stages of the test development cycle, including design, content development, scale development, field testing, administration, equating, validity studies, scoring, and reporting. Leaders will demonstrate how accessibility, as it is woven into test development, must be documented to provide necessary validity evidence and support assessment claims. Finally, leaders will illustrate how threats to accessibility impact the uses and interpretations of test scores, on both individual and group bases. Participants will have the opportunity to develop research questions and designs, applying their knowledge to an accessibility-focused research agenda. Participants will leave the session with an expanded understanding of the decisions that affect students' access to test constructs and how to best guide stakeholders and decision makers toward more accessible tests. Leaders will provide a resource list as part of the session materials.

Session Organizers:

Anne H. Davidson, EdMetric LLC
Stephanie W Cawthon, University of Texas - Austin
Vitaliy Shyyan, Smarter Balanced Assessment Consortium

026. Creating Interactive Applications with R Shiny

Training Session *8:00 to 5:00 pm*

Hilton San Francisco Union Square: Union Square 3 & 4

This session explores the use of the R Shiny package for creating interactive applications. In many testing, commercial, and academic contexts, there is often a need for specialized applications for custom tasks and analyses. Many of the commercially available statistical software programs are offered in a one-size-fits-all format, and thus often lack the flexibility needed across multiple contexts. Shiny is a free, open-source resource that can be used to build custom applications that can be developed and maintained by persons with only a modest level of R programming skill. These apps can be hosted on a webpage or deployed as standalone executable files, and end users of such apps do not need to know any R programming to successfully use them. Using psychometric tasks as examples, we guide participants through building a simple app in Shiny. After teaching the foundations of a Shiny program, we then expand on this simple program to showcase some of the advanced capabilities of Shiny use, including generating reports and building standalone executable programs. Participants should have a moderate level of R programming ability. More advanced R programmers will still benefit from Shiny information that goes beyond "hello world" examples found on Shiny resource pages.

Session Organizers:

Joshua Goodman, NCCPA
John Willse, UNCG
Christopher Runyon, NBME

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

Training Session *8:00 to 5:00 pm*

Hilton San Francisco Union Square: Union Square 5 & 6

The primary aim of the workshop is to provide participants with the necessary practical experience to use cognitive diagnosis models (CDMs) in applied settings. Moreover, it aims to highlight the theoretical underpinnings needed to ground the proper use of CDMs in practice. In this workshop, participants will be introduced to a proportional reasoning (PR) assessment that was developed from scratch using a CDM paradigm. Participants will get a number of opportunities to work with PR assessment-based data. Moreover, they will learn how to use GDINA, an R package developed by the instructors for a series of CDM analyses (e.g., model calibration, evaluation of model appropriateness at item and test levels, Q-matrix validation, differential item functioning evaluation). To ensure that participants understand the proper use of CDMs, the theoretical bases for these analyses will be discussed. The intended audience of the workshop includes anyone interested in CDMs who has some familiarity with item response theory and R programming language. No previous knowledge of CDM is required. By the end of the session, participants are expected to have a basic understanding of the theoretical underpinnings of CDM, as well as the capability to conduct various CDM analyses using the GDINA package.

Presenter:

Wenchao Ma, The University of Alabama

Session Organizer:

Jimmy de la Torre, Hong Kong University

028. Examining the Consequences of Assessment Design and Use Because Assessment Matters

Training Session *10:30 to 12:30 pm*

Hilton San Francisco Union Square: Union Square 19 & 20

Through the use of a mixture of lecture, hands-on exercises, and group work, participants will: • Learn how to systematically examine the consequences of assessment design and use for both classroom and large-scale assessment programs • Identify what are unintended consequences of testing, generate examples of relevant unintended consequences, and create strategies to mitigate them • Understand the connections between unintended consequences, professional standards, and validity • Become familiar with and actively apply two approaches—Integrated Design and Appraisal Framework (Slomp, 2016) and Theory of Action (Bennett, 2010)—for identifying and mitigating unintended consequences of testing and increasing the use of intended ones. An overview of the literature on the consequences of assessment design and use will be provided with illustrations applied to decisions made when developing workplace English communication and collaboration assessment prototypes. Participants will then be guided through two case studies illustrating the application of the IDAF and ToA models to this assessment context. Participants will then work collaboratively on building a plan of action, extrapolated from these frameworks, that they will apply to a third case study. This training session is open to all NCME members.

Presenter:

Maria Elena Oliveri, Educational Testing Service

Session Organizer:

David H Slomp, University of Lethbridge

029. Methodologies and Tools for Think Alouds and Cognitive Laboratories

Training Session *10:30 to 12:30 pm*

Hilton San Francisco Union Square: Union Square 23 & 24

The validity of inferences that can be drawn from performance on items is critical to providing a valid and meaningful test outcome. Establishing the validity of those inferences can be done using a variety of methods; however, each method impacts what inferences can be drawn. Thus, it is important for practitioners and researchers to understand what methods are available, when they should be used, and what tools are available to implement those methods. In alignment with this year's conference theme, this training session offers practitioners and researchers interviewing tools, strategies, and best practices they need to better understand the response processes an item is measuring and how it is operating for a diverse set of test-takers. We will use a combination of interactive lecture format and hands-on practical experience to allow attendees to (a) learn a variety of methodologies for evaluating the affective, behavioral, and cognitive impacts of item design; (b) gain experience in utilizing these methodologies; and (c) explore a tool that facilitates research endeavors. Attendees should bring their own Windows PC laptops to engage in the practical hands-on session. This in-person workshop complements a digital NCME ITEMS module on this topic, which will be available before the workshop.

Presenters:

Andre Alexander Rupp, Educational Testing Service (ETS)
Jacqueline P. Leighton, University of Alberta

Session Organizer:

Blair Lehman, Educational Testing Service

030. Making Measurement Matter: A Practical Guide for Assessment Directors and Test Users

Training Session *1:00 to 5:00 pm*

Hilton San Francisco Union Square: Franciscan Ballroom B

The body of knowledge needed to successfully manage an assessment program is vast, whether the program consists of only one assessment or many. Those charged with managing these programs may come from very diverse backgrounds, and many may not have any formal training in educational measurement. Without adequate support and access to relevant resources and expertise, assessment directors and test users may find themselves in difficult positions in the event of missteps at any of the phases of the program – from item and test development, to test administration, scoring, and reporting of assessment results. The purpose of this workshop is to explore some of the topics that are key for those both new and experienced in these program management roles. Workshop facilitators will address some of the fundamental ideas regarding educational measurement; challenges and possible solutions to administering tests with fidelity; and how to communicate the interpretations of test scales, scores, and sub-scores to the layperson. The workshop will provide planned opportunities for small-group and whole-group discussions of these topics. Workshop takeaways will be new or refreshed knowledge of the basics of educational measurement, and ideas for addressing some of the key challenges for practitioners.

Presenters:

Arthur Thacker, HumRRO
Stephen Sireci, University of Massachusetts Amherst
Vince Verges, Florida Department of Education

Session Organizer:

Stephen Sireci, University of Massachusetts Amherst

031. Experiencing a New Approach to Reliability for Classroom Assessment

Training Session *1:00 to 5:00 pm*

Hilton San Francisco Union Square: Union Square 15 & 16

Though they are fundamental to scientific measurement, it has been difficult to bring the concepts and techniques for determining reliability and validity into the realm of classroom assessment. This training session will be a hands-on introduction to a new approach for enhancing the reliability of assessments developed by classroom teachers. The technique can be used productively by educational researchers, evaluation specialists, test constructors, trainers and social scientists as well. The session will begin with participants experiencing a classroom assessment from the point of view of a student. They will then receive access and orientation to an online assessment information system that will be used to analyze and interpret assessment results. They will learn how the design of the assessment and the interpretation of results exemplifies a new approach to educational information based on practical learning goals and practical learning outcomes. Working at this level of information solves several longstanding problems associated with the determination and interpretation of reliability of classroom assessments. Participants will be given access the information system and its reliability feature to use in their own settings for one year after the session. Participants must bring computers to permit them to access the online assessment information system.

Presenters:

Paul Zachos, ACASE
Jason Brechko, Glens Falls Central School District/ACASE
Panpan Yang, University at Albany

Session Organizer:

Monica De Tuya, ACASE

032. Using Stan for Bayesian Psychometric Modeling

Training Session *1:00 to 5:00 pm*

Hilton San Francisco Union Square: Union Square 17 & 18

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

Presenter:

Manqian Liao, University of Maryland, College Park

Session Organizer:

Yong Luo, Educational Testing Service

033. A Visual Introduction to Computerized Adaptive Testing

Training Session *1:00 to 3:00 pm*

Hilton San Francisco Union Square: Union Square 19 & 20

The training will provide the essential background information on operational computerized adaptive testing (CAT) with an emphasis on CAT components, CAT simulation, Automated test assembly (ATA), and the multi-stage adaptive testing (MST). Besides the traditional presentation through slides, this training consists of hands-on demonstrations of several CAT key concepts and activities through exercises with visual and interactive tools including a CAT simulator, automated test assembler, MST simulator, and other small IRT tools. Practitioners, researchers, and students are invited to participate. A background in IRT and CAT is recommended but not required. Participants should bring their own laptops and item pools, as they will access the tools that were designed to help the participants understand important CAT concepts and tasks and visualize the simulation results. Electronic training materials will be provided via email prior to the conference. Upon completion of the workshop, participants are expected to have 1) a broader picture about CAT; 2) deeper understanding of the fundamental techniques including simulation, ATA, and MST; 3) an understanding of the costs/benefits/trade-offs of linear vs CAT vs MST test designs; 4) appreciation of the visual techniques used to analyze and present results.

Presenter:

David Shin, Pearson

Session Organizer:

Yuehmei Chien, NWEA

034. An Overview of Psychometric Work at Various Testing Organizations

Training Session *1:00 to 5:00 pm*

Hilton San Francisco Union Square: Union Square 22

An overview of the psychometric work routinely done at various testing organizations will be presented in this training session. The training session will focus on the following topics: (1) outline of operational psychometric activities across different testing companies, (2) hands-on activities to review item and test analyses output, (3) hands-on activities to review equating output, and (4) overview of computerized adaptive testing (CAT) and multi-stage testing (MST) and hands-on activities. If time allows, there will be a brief discussion session regarding factors that affect operational psychometric activities in the CAT ad MST environment. We are hoping that through this training session, participants will get a glimpse of the entire operational cycle, as well as gain some understanding of the challenges and practical constraints that psychometricians face at testing organizations. It is targeted toward advanced graduate students who are majoring in psychometrics and seeking a job in a testing organization and new measurement professionals who are interested in an overview of the entire operational testing cycle. Representatives from major testing organizations (e.g., ETS, Pearson, etc.) will present various topics related to processes in an operational cycle.

Presenters:

JongPil Kim, Houghton Mifflin Harcourt Assessment
Jinghua Liu, The Enrollment Management Association
Ye Tong, Pearson
Hanwook (Henry) Yoo, ETS

Session Organizer:

Hyeon-Joo Oh, Educational Testing Service

035. Cognitive Principles for Assessment Research and Test Development

Training Session *1:00 to 3:00 pm*

Hilton San Francisco Union Square: Union Square 23 & 24

The rise of digitally-based assessment has provided many new design options for technology-enhanced items as well as process data as a window into cognitive steps and strategies. Thus, practitioners and researchers in educational assessment need a working knowledge of fundamental cognitive processes, to support both effective design decisions and theoretically-grounded data interpretation. This interactive training session will be led by cognitive psychologists with extensive experience of applying insights from their field to research and practice in educational assessment. The workshop will provide an introduction to three important theoretical frameworks: (1) working memory theory; (2) cognitive load theory; and (3) multimedia theory. We will connect these theories to the latest empirical findings in item design research. You will learn about and experience a collection of methods to study and measure these processes (e.g., think aloud, log file and eye tracking analyses) and identify opportunities for using these in your own research or practice. This session is suitable for assessment professionals and researchers with little or no prior knowledge of cognitive theories and methods. Instructional approaches will include both didactic tutorials on fundamentals of cognitive theories and hands-on experiential demonstrations of key cognitive methodologies. Required materials are provided; please bring a laptop.

Session Organizers:

Marlit Annalena Lindner, IPN - Leibniz Institute for Science and Mathematics Education
Madeleine Keehner, Educational Testing Service (ETS)

036. Using SAS for Monte Carlo Simulation Studies in Item Response Theory

Training Session *3:30 to 5:30 pm*

Hilton San Francisco Union Square: Union Square 19 & 20

Data simulation and Monte Carlo simulation studies are important skills for researchers and practitioners of educational measurement, but there are few resources on the topic. This four-hour workshop presents the basic components of Monte Carlo simulation studies (MCSS). Multiple examples will be illustrated using SAS including simulating total score distribution and item responses using the two-parameter logistic IRT, bi-factor IRT, and hierarchical IRT. Material will be applied in nature with considerable discussion of SAS simulation principles and output. The intended audience includes researchers interested in MCSS applications to measurement models as well as graduate students studying measurement. Comfort with SAS base programing and procedures will be helpful. Participants are encouraged, but not required, to bring their own laptops. The presentation format will include a mix of illustrations, discussion, and hands-on examples. As a result of participating in the workshop, attendees will: 1) Articulate the major considerations of a Monte Carlo simulation study, 2) Identify important SAS procedures and techniques for data simulation, 3) Adapt basic simulation techniques to IRT-specific examples, and 4) Extend examples to more complex models and scenarios.

Session Organizers:

Brian C Leventhal, James Madison University
Allison Ames, University of Arkansas

037. MIRT Graphics using RShiny

Training Session *3:30 to 5:30 pm*

Hilton San Francisco Union Square: Union Square 23 & 24

The purpose of this four-hour workshop is to walk researchers through 15 different MIRT graphics that they can develop using two-dimensional IRT compensatory model item parameters estimated from assessment data. The focus of the workshop will be the RShiny program MIRTGraph. This program consists of a suite of RShiny programs that requires the user to input a file containing their estimated two-dimensional compensatory item parameters and then will give them the option to run 15 different graphics programs. All graphics can be downloaded to include in reports or articles. The MIRTGraph library is broken down into five categories of graphics for: Individual Items, Total Test, Test Information, Two-Dimensional Ability Estimation, and Differential Item Functioning. The goal of this workshop is to provide graphical tools that will yield greater insight into what tests are measuring and inform future test development. The intended audience is testing practitioners who analyze and interpret test results. The session will have an interactive format where the audience can use their laptops and follow along with the presenters and the walk through the various graphics. Attendees will receive a manual, detailing each of the graphics, and a copy of the R Shiny MIRTGraph program.

Presenter:

Qing Xie, ETS

Session Organizer:

Terry Ackerman, University of Iowa

038. NCME Board Meeting #1

4:00 to 7:00 pm

Hilton San Francisco Union Square: Imperial Ballroom B

NCME Board Meeting #1

039. GSIC Social Event

6:30 to 8:00 pm

Hilton San Francisco Union Square: Salon B

GSIC Social Event

040. Sunrise Yoga

6:30 to 7:30 am

Hilton San Francisco Union Square: Plaza A

Sunrise Yoga

041. How to Achieve (or Partially Achieve) Comparability of Scores from Large-Scale Assessments

Coordinated Session *8:15 to 10:15 am*

Hilton San Francisco Union Square: Continental 4

How much and what types of flexibility in assessment content and procedures can be allowed, while still maintaining comparability of scores obtained from large-scale assessments that operate across jurisdictions and student populations? This is the question the National Academy of Education (NAE) set out to answer in its Study on Comparability of Scores from Large-Scale Assessments. This session presents the major findings from eight papers which explore a host of comparability issues that range from examining: (a) the comparability of individual students' scores or aggregated scores, to (b) scores obtained within single or multiple assessment systems, to (c) specific issues about scores obtained for English language learners and students with disabilities. In each interpretive context, the authors discuss comparability issues as well as possible approaches to addressing the information needs and policy concerns of various stakeholders including state-level educational assessment and accountability decisionmakers/leaders/coordinators, consortia members, technical advisors, vendors, and the educational measurement community.

Participants:

- Comparability of Individual Students' Scores (on what is, at face, the same test) *Charles DePascale, National Center for the Improvement of Educational Assessment; Brian Gong, Center for Assessment; Leslie Keng, National Center for the Improvement of Educational Assessment*
- Comparability of Aggregated Group Scores (on what is, at face, the same test) *Scott Marion, National Center for the Improvement of Educational Assessment; Leslie Keng, National Center for the Improvement of Educational Assessment*
- Comparability Issues Within a Single Assessment System *Mark Wilson, University of California, Berkeley; Richard Wolfe, Ontario Institute for Studies in Education of the University of Toronto*
- Comparability Issues Across Different Assessment Systems *Marianne Perie, Measurement in Practice, LLC*
- Comparability Issues Specific to English Learners *Molly Faulkner-Bond, WestEd; James Soland, NWEA*
- Comparability Issues Specific to Students with Disabilities Requiring Accommodations *Maura O'Riordan, University of Massachusetts Amherst; Stephen Sireci, University of Massachusetts Amherst*
- Comparability for Language and Cultural Groups *Kadriye Ercikan, ETS/UBC; Han-Hui Por, Educational Testing Service*
- Interpreting Test Score Comparisons *Randy E Bennett, ETS*

Session Organizer:

Amy Berman, National Academy of Education

Chairs:

Edward Haertel, Stanford University

James Pellegrino, University of Illinois, Chicago

042. Teaching and Learning "Educational Measurement": Defining the Discipline?

Coordinated Session – Panel Discussion *8:15 to 10:15 am*

Hilton San Francisco Union Square: Imperial Ballroom A

What is educational measurement? Is it a subfield of psychometrics or a unique discipline? What is needed to be an expert in educational measurement? The purpose of this proposed interactive panel session is to build upon a recent curriculum review of 135 graduate programs in educational measurement. The fundamental question posed in that review was "What do our curricula and training programs suggest it means to be an educational measurement expert?" In this session, we plan to ask a different question: what should it mean to be an educational measurement expert in the future? To discuss and debate this and other questions, a panel of five mid-career professors have been assembled who are in leadership positions at five universities with prominent graduate programs known for training students who go on to join the NCME community. Our hope is that this could become the start of a larger initiative and conversation among NCME members and the broader field.

Presenters:

Dan Bolt, UW-Madison School of Education

Derek Briggs, University of Colorado Boulder

Andrew Ho, Harvard Graduate School of Education

Won-Chan Lee, University of Iowa

Jennifer Randall, University of Massachusetts Amherst

Session Organizer:

Derek Briggs, University of Colorado Boulder

Chair:

Suzanne Lane, University of Pittsburgh

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

Coordinated Session *8:15 to 10:15 am*

Hilton San Francisco Union Square: Imperial Ballroom B

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

Participants:

Assessing Non-Cognitive Competencies in Medical Licensure *Miguel Paniagua, National Board of Medical Examiners*

The Use of Non-Cognitive Measures as a Component of Licensure and Certification Measures *Kurt F. Geisinger, Buros Center for Testing*

Comingling Cognitive and Non-cognitive Competencies for Credentialing Exams: A Slippery Slope *Chad Buckendahl, ACS Ventures, LLC*

Assessing Non-Cognitive Competencies in Legal Licensure: Lessons from Neighboring Fields *Joanne Kane, National Conference of Bar Examiners*

Organizer:

Joanne Kane, National Conference of Bar Examiners

Discussant:

Patrick Kyllonen, Educational Testing Service

044. Diving into NAEP Process Data to Understand Students' Test Taking Behaviors

Coordinated Session *8:15 to 10:15 am*

Hilton San Francisco Union Square: Yosemite A

Recently, large scale assessments, including the National Assessment of Educational Progress (NAEP), have transitioned to digitally based assessments (DBAs). Logging timing and behavior data on examinees' interactions with items and delivery interface during the test provides a rich data source, called process data, to examine the relationship between students' testing behaviors and performance. This symposium features three separate studies investigating how process data can be used to identify, classify, and explore students' test taking behaviors, using one block of the 2017 NAEP DBA mathematics grade 4 (N=152,500) administered to a nationally representative sample. The first study examines the non-response patterns (i.e., "omit" and "not reached") in process data and estimate time thresholds of non-response category using machine learning techniques with item-, student-characteristics and process data. The second study uses the time students spend on and between item visits to classify visit behaviors and explore potential motivators for visits, using cluster analysis. The third study analyzes the actions students take within each item visit to identify common action sequences and examine how these differ across items using sequence mining techniques. These studies illustrate how research using process data can contribute to discourses on test assembly, test construction, and test validity issues.

Participants:

Revisiting Omit and Not-Reached Scoring Rule using NAEP Process Data. *Nixi Wang, University of Washington; Ruhan Circi, AIR*

Exploring Item Visits in Process Data and Modeling Students' Visit Behaviors and Intentions *Monica Morell, University of Maryland; Ruhan Circi, AIR*

Understanding Students' Problem-Solving Processes via Action Sequence Analyses *Manqian Liao, University of Maryland, College Park; Ruhan Circi, AIR*

Session Organizer:

Ruhan Circi, AIR

Discussant:

Jonathan Weeks, ETS

045. IRTree Models: The Illus-tree-ous and In-tree-guing Response Process Models

Coordinated Session *8:15 to 10:15 am*

Hilton San Francisco Union Square: Yosemite B

IRTrees encompass a multi-dimensional item response theory family of models that explicitly model response processes. These tree-like models have recently garnered significant research attention due to their vast number of diverse applications combined with the wide-spread availability of advanced computing power. The five presentations in this session will exhibit the versatility and utility of the IRTree family of models through empirical applications and methodological innovation research. Presentation 1 will present a novel approach to incorporating response style information from anchoring vignettes in order to estimate a trait of interest from self-report items. Presentation 2 will use an IRTree to account for extreme responses on a topic susceptible to extreme and sensitive opinions. The third presentation will explore the validity of IRTree inferences through factor analysis, simulation, and the calculation of pseudo-item information. The fourth presentation will show the diversity of applications IRTrees can be applied to by estimating an ability when two-attempt multiple choice items are used. Finally, the fifth presentation will assess the stability of response styles by employing an IRTree model across multiple independent traits. The five presentations will jointly display the flexibility and usefulness of the IRTree response process models.

Participants:

A Tree-Based Approach to Identifying Response Styles with Anchoring Vignettes *Brian C Leventhal, James Madison University; Christina Zigler, Duke University School of Medicine*

Using IRTrees Methods to Assess Response Styles in College Student's Abortion Attitudes *Tiffany L Marcantonio, University of Arkansas; Wen-Juo Lo, University of Arkansas; Allison Ames, University of Arkansas; Ronna Turner, University of Arkansas; Kristen Jozkowski, Indiana University*

Validity Evidence for Response Process Models *Allison Ames, University of Arkansas; John Linde, University of Arkansas*

Modeling Two-Attempt Multiple-Choice Items Using IRTrees *Philip J Grosse, University of Pittsburgh; Clement A Stone, University of Pittsburgh*

An IRTree Method to Investigating Stability of Extreme and Midpoint Response Style *Nikole Gregg, James Madison University; Brian C Leventhal, James Madison University*

Session Organizer:

Brian C Leventhal, James Madison University

046. High Definition Detection of Testing Misconduct

Coordinated Session *8:15 to 10:15 am*

Hilton San Francisco Union Square: Yosemite C

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

Presenters:

Huijuan Meng, Graduate Management Admission Council
Danielle Lee, National Council of State Boards of Nursing
yiqin pan, University of Wisconsin-Madison
Mingjia Ma, University of Iowa

Participants:

Impact of RT Model Selection in Detecting Aberrant Test-taking Behaviors *Huijuan Meng, Graduate Management Admission Council*

Hybrid Threshold-Sequential Procedures for the CAT Security *Danielle Lee, National Council of State Boards of Nursing; Hong Qian, National Council of State Boards of Nursing*

A Weak Supervised Learning Approach for Detecting Item Preknowledge in CAT *yiqin pan, University of Wisconsin-Madison; Edison M Choe, Graduate Management Admission Council*

Improving Test Security Analysis Through Noise Removing *Mingjia Ma, University of Iowa; Zhongmin Cui, ACT, Inc.*

king the Longest Identical String Longer *Zhongmin Cui, ACT, Inc.*

Session Organizer:

Zhongmin Cui, ACT, Inc.

Chair:

Zhongmin Cui, ACT, Inc.

Discussant:

James Wollack, University of Wisconsin

047. Assessing Student Learning Outcomes

Paper Session *8:15 to 10:15 am*

Hilton San Francisco Union Square: Continental 2

Participants:

A Longitudinal and Hierarchical Diagnostic Classification Model for Evaluating Learning Progressions *Matthew Madison, Clemson University*

To empirically evaluate learning progressions, this study introduces a model combining a longitudinal diagnostic classification model and the hierarchical diagnostic classification model. This fusion of models allows for the simultaneous examination of attribute hierarchies and student learning over time, which together, comprise the basis of a learning progression.

Exploration of Using Information from Incorrect Options in Reporting Subscores *Hulya Yigit, University of Illinois at Urbana-Champaign; Louis Roussos, Measured Progress; Xi Wang, Measured Progress*

We investigate using information from incorrect response options on multiple-choice items to improve subscore reliability. Option characteristic curves are estimated with B-spline methodology. Pattern scoring using the nominal responses is compared to standard dichotomous scoring methods in both simulation studies and in a real data analysis.

Identifying Online Learning Behaviors Using Log Data from Learning Management Systems *Chang Lu, University of Alberta; Okan Bulut, University of Alberta; Carrie Demmans Epp, University of Alberta; Mark J. Gierl, University of Alberta*

This study investigated students' learning behaviors and academic achievement by extracting features from learning management system log files to develop a predictive model and cluster students into different categories. Results revealed students' common online learning behavior patterns and identified important variables for behavioral classification and performance prediction.

Process-based measurement of listening component skills *Johannes Naumann, University of Wuppertal; Tobias Richter, University of Wuerzburg; Maj-Britt Haffmanns, University of Kassel; Julia Schindler, University of Wuerzburg; Patrick Dahdah, University of Wuppertal; Yvonne Neeb, DIPF - Leibnitz Institute for Research and Information in Education*

A test battery for the assessment of the efficiency of six listening component processes on the word and sentence level is introduced. First results evidence reliability, construct, and criterion validity. Correlations with corresponding reading processes are moderate, pointing to listening as a separate skill, and the importance of its measurement.

Students' Development of Economic Knowledge Over the Course of a Bachelor's Degree *Jasmin Schlax, Research Assistant; Marie-Theres Nagel, Research Assistant; Olga Zlatkin-Troitschanskaia, Johannes Gutenberg-University Mainz; Carla Kühling-Thees, Research Assistant; Judith Jitomirski, Research Assistant; Roland Happ, Research Assistant*

In Germany, a test for measuring economic knowledge was developed and validated, enabling the analysis of knowledge development over the course of studies. We present longitudinal results on the development of knowledge of bachelor students throughout their complete course of bachelor studies based on 4 measurements, each one year apart.

Chair:

William Lorie, Center for Assessment

Discussant:

Paul Nichols, NWEA

048. Cognitive Diagnostic Modeling #1

Paper Session *8:15 to 10:15 am*

Hilton San Francisco Union Square: Continental 3

Participants:

A Deep Learning Approach to Cognitive Diagnostic Modeling: DVAE-CDM *Qi Guo, University of Alberta; Maria Cutumisu, University of Alberta; Ying Cui, University of Alberta; Jacqueline P. Leighton, University of Alberta*

This study proposes to use a discrete variational autoencoder, which is a powerful deep learning model, to estimate a Cognitive Diagnostic Model (CDM). A simulation study showed that the proposed approach performed well in all conditions compared with DINA and DINO, yielding the lowest error rates in most conditions.

A General Nonparametric Classification Method for Polytomous Cognitive Diagnostic Data *Yanhong Bian, Rutgers, the State University of New Jersey; Chia-Yi Chiu, Rutgers, the State University of New Jersey*

This research aims to extend the general nonparametric classification method for binary cognitive diagnostic data to a general method that deals with polytomous attributes and responses. The proposed method was evaluated by intensive simulation studies and results show it allows for small sample sizes to achieve high classification accuracy.

A Nested Diagnostic Classification Model for Multiple-choice Items *Ren Liu, University of California, Merced; Haiyan Liu, University of California, Merced*

The study proposes a general DCM that includes distractor information. An attractive feature of the proposed DCM is that it does not change the psychometric property of the correct response option. In other words, it allows for binary scoring (for correctness) and polytomous scoring (for distractors) at the same time.

Classification Consistency and Accuracy for Cognitive Diagnostic Assessment with Nonparametric Classification Methods *Chia-Yi Chiu, Rutgers, the State University of New Jersey; Yu Wang, Rutgers, The State University of New Jersey*

The study is aimed to develop the classification consistency and accuracy indices for cognitive diagnosis assessments with the general nonparametric classification (GNPC; Chiu, Sun, & Bian, 2018) method. The development allows for the establishment of validity and reliability and thus, improves the usefulness of the assessment.

Identifying subtypes of Mathematical Disability: An approach based on Cognitive Diagnosis Models *Xiangzi OUYANG, The university of Hong Kong; Xiao Zhang, The university of Hong Kong; Tuire Koponen, University of Jyväskylä; Lerkkanen Marja-Kristiina, University of Jyväskylä; Pekka Rasanen, University of Jyväskylä*

The study used cognitive diagnosis models (CDMs) to identify the subtypes of students with Mathematical Learning Disability (MLD) by diagnosing their difficulties in four early numerical skills. We identified seven main subtypes among 99 MLD students and validated the classification results.

Chair:

Anthony Albano, University of California, Davis

Discussant:

Jonathan Templin, University of Iowa

049. New Methodology in Assessments

Paper Session *8:15 to 10:15 am*

Hilton San Francisco Union Square: Franciscan Ballroom A

Participants:

A multiple item response strategy model for educational assessment data *Luo Jinwen, UCLA; Minjeong Jeon, UCLA*

This paper developed a new psychometric modeling approach for educational assessment data which captures differential item response strategies that examinees apply during assessments. The proposed model is based on a finite-mixture IRT model that enable identifications of response strategy switch across subjects and items.

Development of an Alternative Algorithm for the Use of Plausible Values *Andrew Kolstad, P20 Strategies LLC*

In assessment surveys, one normally calculates the standard error of the mean by summarizing the average and variance of each set of PVs. My alternative estimates more stable standard errors by summarizing the average and variance of each examinee's PVs. I evaluate the two methods with simulated and real data.

Multilevel Diagnostic Item Response Modeling for School Diagnostics and Assessment *Meredith Langi, UCLA; Minjeong Jeon, UCLA*

We propose a new item response model for school assessment and diagnostics. The key idea is to apply strategic constraints to the parameters of a multilevel mixture IRT model to capture differences in schools' impacts on student learning outcomes. An application to a large-scale assessment is presented with simulation studies.

Re-envisioning rater training through learning principles: The TALIS Video Study *Courtney Bell, ETS; Katherine Castellano, Educational Testing Service; Yi Qi, ETS; Mariana Barragan Torres, UCLA*

This paper describes a novel approach to the training of raters in the TALIS Video study, an OECD observational study of teaching and learning in eight country/economies. The study conceptualizes rater training as adult learning in a socio-historical context. The paper describes rater certification, calibration, and validation results of training.

The Transformed Gumbel Model (TGM) for Item Responses *Jonathan Darrell Rollins, West Virginia Department of Education; Elena Nightingale, Georgia Department of Education*

A new item response model, the transformed Gumbel model (TGM), models both item complexities and person abilities on a closed interval between zero and one. This research presents properties of the model parameters and resulting probabilities and explores application through simulation studies and real data analyses.

Chair:

Edgar I Sanchez, ACT

Discussant:

Richard Luecht, University of North Carolina

050. Communication of Critical Measurement Issues

Paper Session *8:15 to 10:15 am*

Hilton San Francisco Union Square: Franciscan Ballroom C

Participants:

A framework to improve the utility and defensibility of subscores *Victoria Tamar Tanaka, The University of Georgia; Chris Domaleski, Center for Assessment*

Stakeholders frequently place high value on providing subscores for summative assessments. However, subscores often lack clear meaning or sufficient evidence. The purpose of this study is to provide a framework to help developers improve the clarity and utility of subscores supported by an evidence base tied to the intended interpretations.

Combining Bifactor and Generalizability Theory Models to Enhance Interpretation of Scores *MURAT KILINC, University of Iowa; Walter P. Vispoel, University of Iowa; Mingqin Zhang, University of Iowa; Carrie A. Morris, University of Iowa*

We illustrate two ways of integrating the bifactor and G-theory models to account for multiple sources of measurement error, assuming either congeneric of essential-tau-equivalent relationships among item scores. Results were similar across models with both allowing for partitioning of trait and measurement error variance at score composite and subcomponent levels.

Construct Evidence of a Second Language Listening Test Used for Canadian Immigration *Angel Arias, University of Ottawa; Jean-Guy Blais, University of Montreal*

This study gathers construct evidence for a listening test through confirmatory factor models that were specified following the recommendations of an expert panel with varied backgrounds in applied linguistics. The models fit the data well, providing validity evidence, but suggesting potential construct underrepresentation for one of the test forms examined.

Limitations of research findings in policy decisions: the role of head vs heart *Stanley N Rabinowitz, Pearson; Jon S. Twing, Pearson Assessments*

Researchers envision a world where policy decisions are primarily data driven. Increasingly we confront a skepticism towards research itself, especially when our findings contradict popular practice or long-held beliefs. We describe several assessment innovations across the world that have been affected by this phenomenon, as well as proposed solutions.

Using Admission Test Scores to Improve College Retention Rates *Jessica Patricia Marini, College Board; Paul Westrick, College Board; Emily Shaw, College Board*

This study examines the validity of the new SAT as a predictor of college retention by subgroup. SAT scores showed clear positive relationships with retention across all subgroups. The results inform how colleges can utilize admission test scores to identify students less likely to return to better support their success.

Chair:

Michelle Boyer, Center for Assessment

Discussant:

Catherine Taylor, University of Washington

051. Electronic Board. Saturday 10:35

Electronic Board Session *10:35 to 12:05 pm*

Hilton San Francisco Union Square: Salon A

Participants:

A Diagnostic Classification Models Implementation in Middle School Physics *Kun Su, UNCG; Robert Henson, University of North Carolina at Greensboro*

This study plans to describe and illustrate the process of implementing a Diagnostic Classification Models (DCMs) for a middle school physics test. This process includes defining the attributes, Q-Matrix construction, test development and administration, data collection and analysis, and the intervention. Validity evidence will be collected after the intervention.

A Fixed Anchor Calibration for Equating under Simple Structure Multidimensional IRT *Kazuki Hori, Virginia Tech; Hirotaka Fukuhara, Pearson; Tsuyoshi Yamada, Okayama University*

This study investigates the performance of fixed anchor calibration for equating under multidimensional IRT model. Results from a simulation study show that the fixed calibration can yield substantially biased estimates for item slopes when true factor correlations are different across forms and the correlations of the equated form are fixed.

A Nonparametric Framework for MST with Intersectional Routing for Small Sample *Seongeun Hong, University of Massachusetts Amherst; John Denbleyker, HMH; Scott Monroe, University of Massachusetts Amherst*

In multistage testing (MST), it is important to determine the number of stages for controlling the level of adaptivity. The MST with intersectional routing (ISR) approach improves measurement efficiency and test optimality. We propose an extension of this approach for improving the adaptivity of the test with small samples.

A Variable-Length Stopping Rule for Nonparametric CD-CAT *Chia-Yi Chiu, Rutgers, the State University of New Jersey*

The key rationale of the proposed variable-length stopping rule for the nonparametric CD-CAT is to establish the cutoff by first approximating the probability of a correct item response by the weighted ideal response used in the general nonparametric classification method, which then allows for the estimation of the posterior probabilities.

An Investigation into Calibration and Linking of Integrated Tests *Feifei Li, Educational Testing Service*

Integrated tests measure integrated proficiencies that are presented in complex tasks. Given limited testing time, students cannot be tested on all items but are assigned with item blocks. This study is intended to examine the effects of particular factors in calibrating and linking the multi-dimensional tests with block designs.

DIF Detection in the DINA Model Using the Lagrange Multiplier Test *Yuxiao Zhang, The University of Hong Kong; Yan Sun, Rutgers University; Jimmy de la Torre, Hong Kong University*

This study proposed the Lagrange Multiplier (LM) test to detect DIF in the context of the DINA model. The results show it is more conservative than the Wald test when item quality is medium or low, and performs better as sample size, test length, DIF size or item quality increases.

DIF in MST Routing in the Context of International Large-Scale Assessments *Montserrat Beatriz Valdivia, Doctoral student; Dubravka Svetina Valdivia, Indiana University; Leslie Rutkowski, Indiana University Bloomington*

Through simulation, we examine the effects of differential item functioning (DIF) on routing decisions in the context of multi-stage testing (MST). Additionally, we investigate method's ability to detect DIF in MST in the context of international large-scale assessments. Results suggested presence of DIF diminished the potential benefits of MST.

Effect size measures for differential item functioning in cognitive diagnostic models *Yanan Feng, Indiana University Bloomington*

This study develops an effect size measure of DIF detection for CDMs and explores the blended decision rule that combines the effect size and significance test. A simulation study was conducted to determine if the effect size affect the Type I error and power rates for the Wald procedure.

Examining Cognitive Diagnostic Modeling in Small Sample Contexts *Justin Paulsen, HumRRO; Dubravka Svetina Valdivia, Indiana University*

Cognitive diagnostic models (CDMs) are a family of psychometric models designed to provide categorical classifications for multiple latent attributes. To a large extent, CDMs have been conducted using large scale samples. This study considers the possibility of estimating CDMs in classroom-assessment type conditions using a simulation study.

Hybrid Computerized Adaptive Testing: Combined Mst And Testlet Cat *HsinRo Wei, Riverside Insights; M. David Miller, University of Florida*

This simulation study is combining MST and testlet CAT to investigate how the test length and combinations of MST and testlet CAT, and their interaction may impact to the accurate estimation. Results showed the long test length or more testlets provided better measurement reliability.

Impact of Q-matrix Misspecification on Classification Accuracy of the Nonparametric Classification Methods *Xiaojian Sun, Southwest University; Tao Xin, Beijing Normal University, PRC.; Naiqing Song, Southwest university*

The impact of Q-matrix misspecification on the classification accuracy for nonparametric classification (NPC) and general NPC (GNPC) methods is examined. Results show that: Proportion of item misspecification has negative effect on GNPC method; while it has positive effect on NPC method when entries of the Q-matrix are missing.

Monte Carlo, Post-Hoc and Hybrid Simulations in Multistage Testing *Halil I Sari, Kilis 7 Aralik University; Sakine Gocer Sahin, WIDA at WCER at UW-Madison*

The purpose of the study is to conduct three different simulation studies as monte-carlo, post-hoc and hybrid simulations in multistage testing (MST) environment, and compare the outcomes produced by the three. The aim is to discuss the benefits, advantages and disadvantages of different simulation methods in the MST framework.

Should Interim and Final Scoring Methods Be Consistent in CAT? *Chunxin (Ann) Wang, ACT Inc; Yi He, ACT, Inc.; Jie Li, ACT, Inc.*

This study investigates the effects of consistency in interim and final scoring methods for fixed-length computerized adaptive tests (CAT). Results will provide information on the impact of applying the same or different scoring methods in interim scoring and final scoring on final theta score estimation in practice.

Subdomain Classification for Remediation with a Summative CAT *Xiaodan Tang, University of Illinois at Chicago; Chen Li, Kaplan Test Prep*

Subdomain classification using the subdomain ability and its conditional standard error of measurement (CSEM) may produce non-informative classification, especially when the response sequence is short. This study adopts cognitive diagnostic modeling (CDM) approach to classify subdomain performance in a variable-length CAT to provide remediation suggestions along with the overall evaluation.

The Application of Multiple Group IRT Linking to the Assessment of Test Speededness *Irina Grabovsky, National Board of Medical Examiners*

This project investigates the effect of time pressure on the performance of examinees on a high-stakes examination administered under a time limit. The study establishes a common metric for the comparison of item difficulty in high-stakes tests, allowing for the calibration of tests administered under different timing conditions. Results suggest there is only a small effect of item duration on the test performance of the total group.

The Effect of Anchor Test Composition on Testlet-Based Test Equating *Hongyu Diao, Educational Testing Service; Qianqian Pan, The University of Hong Kong; Junhui Liu, Educational Testing Service*

This study investigated whether the composition of the anchor test can influence the equating on a testlet-based assessment under a NEAT design. The results showed evenly collected items from multiple testlets can produce lower equating bias.

The Effect of Calculator Use on Student Mathematics Assessment Performance *YUAN HONG, American Institutes for Research; Stephan Ahadi, American Institutes for Research*

The online test delivery system allows for calculator use during the mathematics assessments. Different states have different policies on the use of the calculator across grades. This study conducted a series of matched samples comparison and found no significant impact of calculator use on student performance.

Towards a Unified Model Fit Framework: Limited-Information Item Residuals for Diagnostic Classification Models *Daniel Jurich, National Board of Medical Examiners*

The development of limited-information statistics applicable to full-information algorithms has facilitated the assessment of fit in diagnostic classification models. This study demonstrates the potential of bivariate residuals produced when computing these statistics for evaluating item fit. We highlight the comprehensive assessment of model and item fit the limited-information framework enables.

Wellbeing Assessment Validation and Measurement Equivalence Examinations with CFA and IRT *Xinyu Ni, Wake Forest University*

Confirmatory factor analyses (CFA) were used to examine construct validity and measurement equivalence across multiple groups of the Wellbeing Assessment Likert scale (N=11,920). Then two effect size indices calculated with CFA mean and covariance structure (MACS) and Mental-Haenszel test method were compared in evaluating the item-level impact on measurement non-equivalence.

National Council on Measurement - ITEMS Module 1 *Andre Alexander Rupp, Educational Testing Service (ETS)*

Learn about your opportunity to publish in the ITEMS series. The ITEMS portal is your entry point into a world of learning in educational measurement and assessment. ITEMS modules are its centerpiece, which are short self-contained lessons with various supporting resources that facilitate self-guided learning, team-based discussions, as well as professional instruction.

052. Assessment Literacy: Practical Implications and Outcomes

Coordinated Session *10:35 to 12:05 pm*

Hilton San Francisco Union Square: Continental 4

During last year's NCME/National Association of Assessment Directors (NAAD) symposium in Toronto, a panel of noted assessment literacy experts and district-level practitioners addressed the question of whether NAAD should (a) coordinate an inter-organizational initiative to raise assessment literacy levels across a broad range of stakeholder groups or (b) conduct an assessment literacy campaign that focused more narrowly on practicing assessment directors only. After lively discussion, which included considerable input from the audience, a consensus was reached: NAAD should begin narrowly and expand the effort as traction and momentum increase. Accordingly, the presenters in this symposium will either directly address particular aspects of assessment literacy, per se, or more obliquely address topics related tangentially to assessment literacy. Because NAAD's leadership has already decided to continue the initiative into the 2020-2021 schoolyear, this symposium represents, in effect, an interim report. Audience input will again be welcomed.

Presenters:

Leigh Bennett, Loudoun County Public Schools
An-Thy Nguyen Bey, Norfolk Public Schools
Aigner Picou, The Learning Agency
Jeffrey Smith, Township High School District 214
Edward Roeber, MSU

Session Organizer:

Stephen Court, National Association of Assessment Directors

053. Fairness, Equity, and Consequences in College Admissions

Coordinated Session – Panel Discussion *10:35 to 12:05 pm*

Hilton San Francisco Union Square: Imperial Ballroom A

In this coordinated session, a panel of thinkers will engage with one another, and the audience, to discuss the good, the bad, and the ugly of college admissions in the 21st century. Many of us in the field of measurement are here because we believe that high-quality assessments have an important role to play in the college admissions process. When things are working properly, standardized admissions tests have the power to help talented examinees show us what they know and can do, even if their circumstances might otherwise occlude their potential. But is that all these tests do? And do they do that well enough for us to feel like our work is really achieving its intended outcome? And what is our responsibility as measurement professionals, given that we cannot solve pervasive issues of inequality and discrimination that start before and extend beyond the limits of a single standardized test?

Presenters:

Wayne J. Camara, ACT
Jessica Howell, The College Board
Tameka Porter, WCER
Darius Taylor, University of Massachusetts, Amherst

Session Organizer:

Molly Faulkner-Bond, WestEd

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

Coordinated Session *10:35 to 12:05 pm*

Hilton San Francisco Union Square: Imperial Ballroom B

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

Presenters:

Liru Zhang, Delaware Department of Education
Shudong Wang, NWEA
Eric Zibert, California Department of Education

Participants:

Validity and Psychometric Properties of Integrated Item Cluster in Science Assessment *Liru Zhang, Delaware Department of Education*

An Investigation of Efficiency and Validity Evidence in Scoring Innovative Items *Shudong Wang, NWEA; Liru Zhang, Delaware Department of Education*

Examination of Statistical Properties and Accessibility of Technology-Enhanced Items *Eric Zibert, California Department of Education*

Session Organizer:

Liru Zhang, Delaware Department of Education

Discussant:

Richard Patz, University of California-Berkeley

055. Approaches and Considerations to Reducing Test Length of Linear Fixed Form Assessments

Coordinated Session *10:35 to 12:05 pm*

Hilton San Francisco Union Square: Yosemite A

A wide variety of educational assessments are currently being deployed in our school systems. As a result, many state testing clients must deal with ongoing concerns regarding over-burdening students with these assessments. Pressures from various stakeholders, such as parents and school leaders, often focus on the amount of time students spend testing as an option for reducing perceived testing burden. When tasked with reducing the time students spend testing and the perceived testing burden, various assessment programs have explored options to reduce the length of the overall test form (i.e., number of items on a test form). This session provides an overview of perspectives and strategies that have been considered for operational implementation when consulting with state clientele. The results are also applicable to other assessment settings where testing time is a concern.

Participants:

Reviewing Two Test Length Reduction Strategies Implemented by State Assessments *Matthew Gaertner, WestEd*

Benefits of Pattern Scoring with an Abbreviated Assessment *Jennifer Beimers, Pearson; Joseph Fitzpatrick, Pearson; Jasmine Carey, Colorado Department of Education; Joyce Zurkowski, Colorado Department of Education*

Evaluation of Measurement Precision on Shortened Forms *Yang Lu, ACT, Inc.*

Evaluating Cutscores and Classification Accuracy with Reduced Test Lengths *Eric Moyer, Pearson; Samuel Haring, ACT*

Session Organizer:

Samuel Haring, ACT

Discussant:

Joanne Jensen, WestEd

056. Challenges with Automatic Item Generation Implementation: Recent Research, Strategies, and Lessons Learned

Coordinated Session *10:35 to 12:05 pm*

Hilton San Francisco Union Square: Yosemite B

This session is important for organizations considering AIG implementation, as well as organizations currently implementing AIG, as these papers address issues that an organization is likely to encounter with AIG implementation. Over the past decade, the educational measurement field has seen great enthusiasm for automatic item generation (AIG). Producing more items in a faster time frame without increasing costs is a priority in many organizations, which is where AIG can be helpful. AIG protects test security by allowing for exposed items to be used less frequently, in turn protecting the validity of assessment scores. However, committing to AIG implementation is initially a significant cost investment. In order for AIG to deliver on its potential for improving assessment development processes, more studies are needed to demonstrate its effectiveness. Examples of successful AIG implementation in the literature, and discussions of lessons learned while facing common challenges during AIG implementation, would greatly benefit the educational measurement community. The focus of this coordinated session is recent research and strategies to inform successful AIG implementation. These papers from four unique organizations share the common theme of addressing challenges that are encountered when implementing AIG and is valuable for attendees from organizations interested in AIG implementation.

Participants:

Operationalizing AIG: Talent, Process, and System Implications. *Donna Matovinovic, ACT Inc.*

Items and Item Models: AIG Traditional and Modern *Emily Bo, NWEA; Wei He, NWEA; Abby Javurek, NWEA; Sarah Miller, NWEA; Sylvia Scheuring, NWEA; Naimish Shah, NWEA; Mary Ann Simpson, NWEA*

Exploring the Utility of Semantic Similarity Indices for Automatic Item Generation *Pamela Kaliski, American Board of Internal Medicine; Jerome Clauser, ABIM; Matthew Burke, ABIM*

The Effect of Using AIG Items to Pre-equate Forms. *Drew Dallas, NCCPA; Joshua Goodman, NCCPA*

Session Organizer:

Pamela Kaliski, American Board of Internal Medicine

Discussant:

Hollis Lai, University of Alberta

057. Development and Empirical Recovery of a Learning Progression that Incorporates Student Voice

Coordinated Session *10:35 to 12:05 pm*

Hilton San Francisco Union Square: Yosemite C

In keeping with the conference theme, Making Measurement Matter, we will discuss research work on building and validating a learning progression-based assessment for the concept of function, a keystone in students' understanding of higher mathematics. This effort also speaks to the goals of "bringing outside voices into the educational measurement community," and fairness as equal priorities. The research includes students and schools served by research collaborators seeking to improve mathematics education for students characterized as underserved in mathematics. Recently, we conducted a computer-delivered pilot of tasks in which data from 1102 students were collected. The first two speakers will focus on the theory and design behind the assessment. The first will discuss the overall design of the project, and summarize work conducted to date. The second will discuss how student feedback on the tasks was elicited during the focus groups, with the intent of making revisions that would enhance meaningfulness and clarity. The last two speakers will discuss outcomes from the pilot, discussing student responses and what they suggest about the validity of the LP, and psychometric results concerning the empirical recovery of the levels of the LP, an essential step in its validation.

Participants:

Steps in the Design and Validation of the Assessment: An Overview *Edith Aurora Graf, Educational Testing Service; Maisha Moses, Young People's Project; Cheryl Eames, Southern Illinois University Edwardsville; Pater van Rijn, ETS*

Eliciting Student Feedback on the Assessment Through Focus Groups *Maisha Moses, Young People's Project*

Response Analysis using the Finite-to-Finite Strand of the Learning Progression for the Function Concept *Cheryl Eames, Southern Illinois University Edwardsville*

Psychometric results for two strands of a learning progression for the concept of function *Pater van Rijn, ETS; Edith Aurora Graf, Educational Testing Service*

Session Organizer:

Edith Aurora Graf, Educational Testing Service

Discussant:

Frank E Davis, Frank E. Davis Consulting

058. Scrutinizing Item Responses and Response Times: Experimental and Analytic Approaches

Coordinated Session *10:35 to 12:05 pm*

Hilton San Francisco Union Square: Continental 1

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

Participants:

Measurement invariance across different scoring and timing conditions in joint modeling of item responses and response times *Usama Ali, ETS; Peter van Rijn, ETS Global*

Effect of immediate feedback on performance in practice tests *Yigal Attali, ETS; Usama Ali, ETS*

Impacts of item types in the response-time conditional dependencies *Hyo Jeong Shin, educational testing service; Paul A Jewsbury, ETS*

Time-accuracy conditional dependencies and latent regression models *Paul A Jewsbury, ETS; Hyo Jeong Shin, educational testing service; Peter van Rijn, ETS Global*

Session Organizer:

Usama Ali, ETS

Chair:

Peter van Rijn, ETS Global

Discussant:

Dylan Molenaar, University of Amsterdam

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

Coordinated Session *10:35 to 12:05 pm*

Hilton San Francisco Union Square: Continental 2

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

Participants:

Data and methodological strategies for analyzing response process data in practice *Jiangang Hao, Educational Testing Service; Chen Li, Educational Testing Service; Robert J Mislevy, Educational Testing Service*

Exploring the progression of writing fluency in large-scale assessments using keystroke logs *Yang Jiang, Educational Testing Service; Jiangang Hao, Educational Testing Service*

Assessment Platform Tool Usage Analytics *Jie Gao, Educational Testing Service*

Sequence Analytics to Understand Item Transitions during Assessment *Mengxiao Zhu, Educational Testing Service*

Response time and its relationships with students and testing characteristics *Bingchen Liu, ETS*

Session Organizer:

Jiangang Hao, Educational Testing Service

060. Advances in IRT

Paper Session *10:35 to 12:05 pm*

Hilton San Francisco Union Square: Continental 3

Participants:

A Latent Dirichlet Allocation Model of Action Patterns *Dakota Cintron, University of Connecticut; Xiang Liu, Educational Testing Service; Qiwei He, Educational Testing Service*

This research uses latent Dirichlet allocation topic models on action pattern data from a large-scale assessment. With these models, test taker action patterns for an item are modeled as a document of words. This research demonstrates how problem-solving strategies can be derived from topic distributions of action pattern data.

Extending G-theory Models to Account for Scale Coarseness and Congeneric Relationships *Walter P. Vispoel, University of Iowa; Guanlan Xu, University of Iowa; Wei S Schneider, University of Iowa; Mingqin Zhang, University of Iowa; MURAT KILINC, University of Iowa; Carrie A. Morris, University of Iowa*

We used confirmatory-factor-analysis-based, G-theory models to account for multiple sources of measurement error, scale coarseness effects, and congeneric relationships among item scores. Results for G-coefficients revealed few differences between congeneric and essential-tau-equivalent models, but consistent differences favoring continuous-latent-response-variable over raw-score models with scales having from two to eight response options.

Multilevel Reliability Measures of Latent Scores within an Item Response Theory Framework *Sun-Joo University Cho, Vanderbilt University; Jianhong Shen, NA; Matthew David Naveiras, Vanderbilt University*

This paper evaluated multilevel reliability measures in two-level nested designs (e.g., students nested within teachers) within a multilevel item response theory (IRT) framework. A simulation study was implemented to investigate the behavior of the multilevel reliability measures and the uncertainty associated with the measures in various multilevel designs.

Using Neural Network for Latent Transition Hierarchical Cognitive Diagnostic Models *Chi Chang, Michigan State University*

Modeling longitudinal CDMs with hierarchical attributes in classroom settings provides information about the growth of skill mastery across multiple occasions as well as the effects of interventions that may have happened within the study time span. This paper uses simulations to address the highlights and issues of the model.

Chair:

James Burns Olsen, Renaissance

Discussant:

Leah Feuerstahler, Fordham University

061. Standard Setting

Paper Session *10:35 to 12:05 pm*

Hilton San Francisco Union Square: Franciscan Ballroom A

Participants:

Improvements to the Bookmark and Item Descriptor Matching Standard Setting Methods *Deborah Schnipke, ACS Ventures, LLC; Russell Keglovits, ACS Ventures, LLC*

The Bookmark and Item Descriptor Matching methods assume panelists will be able to detect patterns in their ratings so they can determine cut scores. When clear patterns do not exist, additional guidance is needed to aid the panelists and/or new analyses are needed. This paper discusses both types of improvements.

Adaptation of the Bookmark Method to Establish Growth Standards *Leslie Vanessa Rosales, Juarez & Associates*

This research presents an evaluation performed to an adaptation of the Bookmark Method conducted to establish new reading growth standards in Guatemala. Evaluation considered different validity sources of information. Results suggest that the adaptation could be used to establish growth standards and that consequential validity should be further explored.

Evidence for Angoff developed passing score criteria with ROC analyses *Kari Hodge, NACE International Institute; R. Noah Padgett, Baylor University; Shanna Attai, Baylor University*

The Angoff method is widely used to establish defensible pass/fail scores for multiple choice exams. Additional validity evidence for cut-scores can be gained with receiver-operating-characteristic (ROC) analysis. Using data from a credentialing program, results from ROC analyses indicated that cut-scores set by Angoff may be too low, implications are discussed

Investigating the Panel Effect using Generalizability Theory under Angoff Standard Setting Context *Seohong Pak, National Board of Medical Examiners*

This research is designed to investigate the impact of panel effect on the credibility of a cut score within Generalizability theory using two different G study designs. The D study is also sought to propose beneficial guidance needed to have highly defensible and stable cut scores.

Optimal Response Probabilities in Embedded Standard Setting *Robert Cook, ACT, Inc; Daniel Lewis, Creative Measurement Solutions*

Many standard setting approaches require consideration of response probability (RP) to establish cut scores. Embedded Standard Setting (ESS) derives cut scores empirically, allowing for empirical selection of an optimal RP. This paper describes the method, benefits, and theory supporting optimal RP selection within, and possibly beyond, the ESS paradigm.

Chair:

Matthew Swain, HumRRO

Discussant:

Karla Egan, EdMetric

062. STEM Learning and Assessment

Paper Session *10:35 to 12:05 pm*

Hilton San Francisco Union Square: Franciscan Ballroom C

Participants:

Development and Validation of a Situational Judgement Test Measuring High School Students' Mathematics Attitudes *Kalina Gjicali, The Graduate Center, CUNY; Anastasiya Lipnevich, The Graduate Center, CUNY and Queens College; Stefan Krumm, Freie Universität Berlin*

The goal of this study was to demonstrate the development of a self-report mathematics attitude measure using innovative item types (i.e., situational judgement tests (SJTs)) and examine the psychometric properties of the Mathematics Attitude Situational Judgement Test (MA-SJT).

Introducing the STEM Teacher Observation Protocol (STEM TOP): Evidence of Internal Structure *Elizabeth Adams, Southern Methodist University; Anthony Sparks, Southern Methodist University; Leanne R. Ketterlin-Geller, Southern Methodist University*

The STEM Teacher Observation Protocol (STEM TOP) was developed to support instructional leaders in coaching middle school science teachers. A multi-level exploratory factor analysis supports that internal structure of the STEM TOP is consistent with measuring teachers' enacted STEM instruction specific to: (a) active learning, and (b) management and discipline.

IRT modeling of decomposed student learning patterns in higher education physics *William B. Walstad, University of Nebraska; Susanne Schmidt, Johannes Gutenberg University Mainz; Olga Zlatkin-Troitschanskaia, Johannes Gutenberg-University Mainz*

This study describes a new sophisticated modeling approach for differentiated analyses of aggregated test scores from a multiple-choice test in physics using students' disaggregated response patterns from the pretest and posttest, which allows for a more precise measuring of change in student learning and understanding over the course of studies.

Isolating Construct-irrelevant Variance in Interactive Simulations for Science Assessment *Jinnie Choi, Pearson; Kristen DiCerbo, Pearson; Emily Lai, Pearson; Matthew Ventura, Pearson*

In technology-rich assessment environments, learners with high-level proficiency in an intended construct may show low performance if they cannot navigate and use the interface properly. This paper describes our approach to isolating construct-irrelevant variance in interactive science simulations, through evidence-centered design, modeling of 'interface use', and examination of validity evidence.

Successes and Challenges of Implementing Individualized Practices in an Online One-on-One Mathematics Tutoring Context *qinjun wang, TAL Education Group; Liwei Wei, TAL Education Group; Yuan Zhao D'Antilio, TAL Education Group; Rui Li, TAL Education Group; Chengfeng Wu, TAL Education Group; Angela Bao, TAL Education Group*

The study investigates the effectiveness of an in-class adaptive practice design based on a learner's cognitive skill performance. Positive results of the learning achievement from treatment group was found, opportunities and potential challenges in implementing individualized practices in Chinese classroom context are discussed.

Chair:

Zachary Feldberg, University of Georgia

Discussant:

Howard Everson, City University of New York

064. Electronic Board. Saturday 12:25

Electronic Board Session *12:25 to 1:55 pm*

Hilton San Francisco Union Square: Salon A

Participants:

A Comparison of Methods for Deriving Composite Scores for Mixed-format Tests *Xia Mao, NBOME*

The study compares four different methods to derive composite scores for tests with mixed formats. The impact of the methods on reliability and decision consistency of the scores is evaluated by using empirical data from a high-stakes medical licensure examination, and the implications for practice are discussed.

A General Cognitive Diagnosis Model for Polytomous Attributes *XUE-LAN QIU, The University of Hong Kong; Jimmy de la Torre, Hong Kong University; Kevin Carl Santos, University of the Philippines; James Zhang, University of Groningen, Netherlands*

Existing cognitive diagnosis models (CDMs) that can accommodate polytomous attributes contain a number of simplifying assumptions. This study proposes a general framework for polytomous-attribute CDMs (pCDMs) with more relaxed assumptions. The necessity of the pCDMs and parameter recovery of the proposed models are shown in two simulation studies.

Analysis of a Statewide Standardized Assessment Data using Multilevel Redundancy Analysis *Kwanghee Jung, Texas Tech University; Hongwook Suh, Nebraska Department of Education; Jaehoon Lee, Texas Tech University; Jungkyu Park, Kyungpook National University; Kyungtae Kim, Tennessee Department of Education*

Multilevel redundancy analysis (MLRA) using dimension reduction features can optimally handle hierarchically structured large-scale data with large number of predictors and criterion variables measured at many different levels. This study illustrates an application of MLRA for educational research, especially its use with a statewide standardized test data.

Comparing Unfolding Medical Case Studies to Their Individual Single-Item Counterparts *William Muntean, Pearson VUE; Joe Betts, NCSBN; Shu-chuan Kao, Pearson; Hong Qian, National Council of State Boards of Nursing*

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

Continuity of Students' Disengaged Responding in Large-Scale Assessments: Evidence from Response Times *H. Cigdem Yavuz, Cukurova University*

This paper attempts to investigate the number of students who continue their disengaged responding behaviors in large-scale assessments. The results reveal that most students continued disengaged responding throughout the test and additional survey items. Such continuous disengaged responding appears to affect parameter estimation negatively. Further implications for practice are discussed.

Data Mining Classification of Mathematics Self-efficacy in Large-scale Assessment *Ya Zhang, Western Michigan University*

This study compared the effectiveness of different data mining techniques in classifying students' mathematics self-efficacy measured on a large-scale assessment (PISA 2012). The student and school level predictors of mathematics self-efficacy were identified. It was found that the multilevel k-means clustering generated superior classifications of the mathematics self-efficacy.

Decomposition of General Ability and Construct-Irrelevant Variance of Online Item Format: Multidimensional IRT Approach *Daeryong Seo, Pearson; Insu Paek, Florida State University; Se-Kang Kim, Fordham Univeristy*

General math ability and construct-irrelevant variance (CIV) were decomposed using multidimensional bifactor model which estimated only item format factors and general ability factor; there was no correlations among the factors. Both gender and ethnicity variables explained the ability as well as CIV caused by online item format.

Equating Words-Correct-Per-Minute (WCPM) Scores in an Oral Reading Fluency Assessment *Jing Chen, Northwest Evaluation Association; Mary Ann Simpson, NWEA*

Words-correct-per-minute (WCPM) scores from dissimilar passages are problematic for monitoring students' reading progress. We develop a method based on graph theory to identify pairwise relationship between passages for equating. Such equated scores provide a better indication of students' oral reading fluency by accounting for differences in passage difficulty.

Evaluate Item Response Similarity among a Group of Examinees *Hongling Wang, ACT, Inc.; Chi-Yu Huang, ACT, Inc.*

This study is to investigate what indices can be used to measure item response similarity among a group of examinees. Those indices can be used to identify the collusion groups that have strong item response similarity. The result of this study has practical implications for test security practitioners.

Four Decades of Measuring Attitude towards Mathematics *Erika Majoros, University of Gothenburg; Monica Rosén, University of Gothenburg; Jan-Eric Gustafsson, University of Gothenburg*

The aim of the present study is to evaluate the feasibility of linking three mathematics attitude scales in international large-scale assessments on mathematics from 1980 to 2015. The purpose is to provide a basis for investigating long-term trends in non-cognitive educational outcomes. Preliminary results show great possibilities for linking.

Graphic exploratory goodness-of-fit tools based on the gradient function for detecting population heterogeneity in IRT models *Jung Yeon Park, KU Leuven; Yoon Soo Park, University of Illinois at Chicago*

This study examines the use of graphical tools based on the gradient function approach to evaluate distributional assumptions for IRT model parameters. Simulation studies were conducted based on item characteristics of psychiatric measures, demonstrating flexible and practical utility of the gradient function method.

Harnessing Electronic Data to Optimize Scale-Up of Assessment Use in Early Intervention *Sondra Marie Stegenga, University of Utah; Daniel Anderson, University of Oregon*

Electronic assessment data sources offer opportunities for quality improvement and scale-up. This paper highlights use of electronic data to examine strategies for improving assessment uptake through use of interrupted time series design with follow-up interviews. This unique mixed methods approach holds potential to improve equity and understanding in assessment use.

Investigating the Impact of Response Noise on Forced-Choice Item Response Theory Models *Bea Margarita Ladaga, University of the*

Philippines Diliman; Joemari Escalona Olea, University of the Philippines Diliman

This study investigates the impact of respondents haphazardly ranking middle options for RANK format forced-choice items. Results reveal that inducing response noise to middle rankings diminishes the performance of the RANK model in terms of person parameter recovery, resulting in MOLE format producing more reliable estimates.

Multilevel IRT Modelling of Reading Development in German Secondary School *Theresa Rohm, Research Associate*

Students are separated early into different school tracks in the German school system. Multilevel growth modelling revealed, that this early segregation fosters social segregation in reading development. Using the alignment method and multilevel structural equation modelling, it was investigated if the results are sensitive to measurement non-invariance across school types.

Parameterizations Matter in Bifactor Models *Wenya Chen, Loyola University Chicago; Ken Fujimoto, Loyola University Chicago*

The scale of the metric underlying a bifactor model are commonly set in two ways. These two parameterizations, in theory, lead to equivalent models. However, we show using real and simulated data that sample size affects whether the theoretical equivalence between the two parameterizations holds when data are analyzed.

Person-fit Indices to Detect Social Desirability Bias Using Dichotomous Items *Sanaz Nazari, University of Florida; A. Corinne Huggins-Manley, University of Florida; Walter Leite, University of Florida*

Social desirability bias (SDB) introduces measurement error in educational assessment. The study goal is to compare person-fit indices to identify SDB according to receiver operating curve (ROC) and model fit evaluation. It is expected that HT shows the largest area under the curve, and IZ* indicates the largest fit improvement.

Predicting Summative Assessment Results from Interim Assessment Performance Using Machine Learning *Luciana Cancado, Curriculum Associates*

This study evaluates various machine learning (ML) algorithms for predicting proficiency on a state summative assessment using interim assessment data and student- and school-level characteristics. Prediction accuracy results of ML algorithms using different sets of predictors indicate that simple classifiers can achieve optimal classification performance under some conditions.

Using Trait Indicators to Measure Growth on a Scenario-Based Assessment *Paul Deane, ETS; Pater van Rijn, ETS; Hongwen Guo, Educational Testing Service; Chen Li, Educational Testing Service; Mo Zhang, ETS*

A scenario-based assessment (SBA) is a standardized assessment designed to model a sequence of component classroom tasks. This paper demonstrates that an English Language Arts SBA can be used as a multidimensional measure, combining task scores, automated writing evaluation features, and writing process features to identify differential school growth patterns after an intervention.

Writing Prompt Calibration with Respect to Local Independence *Dong-In Kim, DRC; Christie Plackner, DRC; Marc Julian, DRC; Vince Struthers, DRC; Mayuko Simon, DRC*

When writing prompts with multiple traits scores are calibrated, local independence assumption can be violated due to a rater's tendency to let their impression of the first trait influence the scoring of other traits. This study examined several different calibration approaches to address the violation of the local independence assumption.

Supplementary resources for communicating assessment results to parents from underserved groups *Priya Kannan, Educational Testing Service; Delano Hebert, Educational Testing Service; Shiyi Shao, Educational Testing Service; Florencia L Tolentino, Educational Testing Service; Ryan M Steinly, University of Pittsburgh; E Caroline Wylie, Educational Testing Service*

Parents sometimes struggle to understand results presented in their child's summative assessment score reports. Score reports and supplementary resources developed in Spanish and English were evaluated with parents from two underserved groups (ELL and low-SES). We will discuss best practices across states and results from interviews with 20 parents.

065. **Invited Session Ed Tech and Measurement: Signaling Skills for the Future of Learning and the Future of Work**

Coordinated Session – Panel Discussion *12:25 to 1:55 pm*

Hilton San Francisco Union Square: Continental 4

The pathways for traversing the education to workforce ecosystem are becoming more complex. A constant throughout this is the need for clear, strong, and valid signals of the skills desired by educators and employers, and those skills attained by learners and workers. Which skills signals do we need now? In 5 years? How strong must these signals be? What is Ed Measurement's role in ensuring that Ed Tech can produce the quality signal. During this session experts will discuss how the ecosystem can better organize itself around our evolving understanding of these skills, as well as our need to develop the capabilities to signal them.

Presenters:

Pat Leonard, Credly
Debbie Durrence, Gwinnett County Public Schools
William G. Harris, Association of Test Publishers

Session Organizers:

Javarro Russell, ETS

066. Digital Activities for Workplace Communication and Collaboration: Opportunities, Frameworks, and Challenges

Coordinated Session *12:25 to 1:55 pm*

Hilton San Francisco Union Square: Imperial Ballroom A

The assessment of complex constructs of communication and collaboration requires close alignment between workplace needs, instruction, and associated professional development resources. These resources can include a variety of digitally-delivered activities that teach learners how to recognize, develop, and use these skills. Multidisciplinary collaborations between experts from different professional fields (e.g., cognitive and learning sciences, assessment design, psychometric modeling, human-computer interaction, and educational technology) are needed to maximize the desired, intended consequences from the use of these activities and minimize undesirable, unintended effects. This coordinated session provides an overview of challenges and opportunities related to the design, implementation, and evaluation of authentic, digitally-delivered activities of English communication and collaboration skills in the workplace for professional development purposes. Throughout the session, we discuss the value of the integration of these kinds of activities into formative assessments focusing on communication and collaboration skills at work. Moreover, we discuss from multiple perspectives the types of methodological challenges that are typically encountered during the assessment design and development process and present various conceptual frameworks that can help interdisciplinary teams conduct this work effectively.

Presenters:

Maria Elena Oliveri, Educational Testing Service
Andre Alexander Rupp, Educational Testing Service (ETS)
David H Slomp, University of Lethbridge
Diego Zapata, ETS

Participants:

Evidence Synthesis for Theories of Action *Maria Elena Oliveri, Educational Testing Service*

The Integrated Design and Appraisal Framework *David H Slomp, University of Lethbridge*

Sociocognitive Considerations for Psychometric Modeling *Andre Alexander Rupp, Educational Testing Service (ETS); Robert J Mislevy, Educational Testing Service; Maria Elena Oliveri, Educational Testing Service*

Effective Communication of Diagnostic Information *Diego Zapata, ETS*

Session Organizer:

Maria Elena Oliveri, Educational Testing Service

Discussants:

Bruno D. Zumbo, UBC
Kurt F. Geisinger, Buros Center for Testing

067. Research for Practical Issues and Solutions in Computerized Multistage Testing (Book by C&H)

Coordinated Session *12:25 to 1:55 pm*

Hilton San Francisco Union Square: Imperial Ballroom B

Since 2014, computerized multistage testing (MST) operational applications have been increasing rapidly, given MST practicability. Meanwhile, researchers have continued to explore approaches to address practical issues and develop software to support the operational applications. With the increasing number of MST operational applications, many testing institutions and researchers have also gained experience in dealing with their operational challenges and solving their practical problems in the context of their large-scale assessments. This symposium presents the most recent research for various practical issues and considerations, methodological approaches, and solutions for implementing MST for operational applications. It is based on an upcoming volume Research for Practical Issues and Solutions in Computerized Multistage Testing (2020) by Chapman and Hall, the sequel volume to the popular volume Computerized Multistage Testing: Theory and Applications (2014) by Chapman and Hall. We expect this symposium to be a great attraction at the conference for practitioners, professors teaching computerized adaptive and multistage testing, and students learning about multistage testing.

Participants:

Purposeful Design for Useful Tests: Considerations of Choice in Multistage-Adaptive Testing *April Zenisky, University of Massachusetts Amherst; Stephen Sireci, University of Massachusetts Amherst*

MST Strategic Assembly and Implementation *Richard Luecht, University of North Carolina; Xiao Luo, Measured Progress*

MST Design and Analysis Considerations for Item Calibration *Paul A Jewsbury, ETS; Pater van Rijn, ETS*

Predicting and Evaluating the Performance of Multistage Tests *Tim Davey, ETS*

Session Organizer:

Duanli Yan, ETS

068. Psychometric Modeling of Item Response Data Based on a Table of Specifications

Coordinated Session *12:25 to 1:55 pm*

Hilton San Francisco Union Square: Yosemite A

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

Participants:

The Role of Test Specifications in the Exam Lifecycle *Marc Kroopnick, Association of American Medical Colleges; Ying Jin, Association of American Medical Colleges*

Extended Multivariate Generalizability Theory *Robert Brennan, Professor Emeritus, University of Iowa, Center for Advanced Studies in Measurement and Assessment*

Using XMGT to Model Changes to a Table of Specifications *Jaime Malatesta, Center for Advanced Studies in Measurement and Assessment*

IRT Approaches to Modeling a Table of Specifications *Won-Chan Lee, University of Iowa; Stella Yun Kim, UNC Charlotte*

Session Organizer:

Jaime Malatesta, Center for Advanced Studies in Measurement and Assessment

Discussant:

Michael Kolen, University of Iowa

069. Identifying Rushing in CAT and Investigating the Effects on Differentiated Instruction

Coordinated Session *12:25 to 1:55 pm*

Hilton San Francisco Union Square: Yosemite B

To attain test scores that validly indicate what a student knows and can do, students must exhibit motivated and effortful behavior throughout the testing event. Therefore, it is pivotal to accurately identify non-effortful behavior. Computer based testing has made possible the use of response time-based measures for separating effortful from non-effortful behavior. This symposium conceptualizes non-effortful behavior as rapid-guessing, e.g. rushing, at the item and test level. The four papers compare different methods for identifying rapid-guessing, introduce frameworks for setting rapid-guessing thresholds at the item and test level, and address the practical implication of using test scores with a high proportion of non-effortful responses.

Participants:

- Developing Item-Level Rush Flags for the i-Ready Diagnostic Assessment *Logan Rome, Curriculum Associates; Elizabeth Patton, Curriculum Associates*
- Developing Test-Level Rush Flags for the i-Ready Diagnostic Assessment *Elizabeth Patton, Curriculum Associates; Logan Rome, Curriculum Associates*
- Evaluating Practical Implications of Using Rushed i-Ready Diagnostic Scores for Instructional Purposes *Lexi Lay, University of North Carolina at Greensboro; Elizabeth Patton, Curriculum Associates; Logan Rome, Curriculum Associates*
- Comparison of Three Methods for Detecting Test-Level Rushing in the i-Ready Diagnostic Assessment *Can Shao, Curriculum Associates; Logan Rome, Curriculum Associates; Yawei Shen, The University of Georgia*

Session Organizer:

Elizabeth Patton, Curriculum Associates

070. Human Rater Thinking and Decision Making

Coordinated Session *12:25 to 1:55 pm*

Hilton San Francisco Union Square: Yosemite C

This session of four papers summarizes applied research that seeks to better understand and ultimately improve the human scoring enterprise by examining how raters think and how those thoughts influence the scores that they assign. The first paper, written by McGrane and Wolfe, compares how raters think when engaged in comparative judgments (e.g., "which essay is better?") versus a categorization task (e.g., "which score level should be assigned?"). Similarly, the second paper, written by Benton, attempts to determine how comparative judgments and categorization decisions compare when the same number of decisions are made about each response. The third paper (by Wolfe & Lockwood) seeks to develop rater-specific models that capture essay features that contribute to the difficulty raters have assigning accurate scores. The fourth paper (by Finn, Morehead, & Dunlosky) focuses on how different strategies for training raters influences the effectiveness of rater training. Jo-Anne Baird of Oxford University will serve as the Discussant for the session.

Participants:

- Examining the category selection versus comparative judgement approaches to writing assessment *Joshua McGrane, Oxford University; Edward W Wolfe, ETS*
- Why does scoring using comparative judgement work so well? *Tom Benton, Cambridge Assessment*
- Rater-Specific Difficult to Score Models for Essays *Edward W Wolfe, ETS; J.R. Lockwood, ETS*
- Using Cognitive Theory to Inform Human Constructed Response Scoring Training *Brigdid Finn, ETS; Kayla Morehead, Kent State University; John Dunlosky, Kent State University*

Session Organizer:

Edward W Wolfe, ETS

Discussant:

Jo-Anne Baird, Oxford University

071. Feasibility of an Embedded Field-Test Model in the Enhancement of Preequating

Coordinated Session *12:25 to 1:55 pm*

Hilton San Francisco Union Square: Continental 1

There is often a push for testing programs that use classical post-equating to move to IRT pre-equating with embedded field-testing for faster reporting and the ability to construct new forms without needing to consider equating designs. This situation is complicated when the tests in question are mixed-format, vertically scaled, and/or are administered in both paper and CBT modes. This coordinated paper session provides a framework for making that switch, detailing the studies and analyses that were conducted, as well as rigorous impact investigations. We discuss issues such as context effects related to embedded field-testing on both filed-test and operational items across item types and final test forms, effects on subgroups of examinees, and effects on individual and aggregate reported scores and level classifications. Final results indicate it is possible to have comparable scores across the change in methodology for all subjects and grades studied if careful development and analyses procedures are followed. In additional to presenting a case study other test developers and researchers can follow, this set of papers provides a broader framework for testing programs wanting to investigate the impact of a methodological change in test equating prior to implementation.

Participants:

An Overview of the Embedded Field-Test Model *Wei Tao, ACT, Inc.*

Evaluating Impacts on Operational Item Performance in the Embedded Field-Test Model *Troy Chen, ACT, Inc.*

Comparing CTT Post-equating and IRT Pre-equating in the Embedded Field-Test Model *Yi-Fang Wu, ACT, Inc.*

Investigating Potential Confounding Testing Mode Effects in Transition from Post-equating to Pre-equating *Yong He, ACT*

Session Organizer:

Yi-Fang Wu, ACT, Inc.

Chair:

Yi-Fang Wu, ACT, Inc.

Discussant:

Deborah Harris, University of Iowa

072. International Assessment Programs

Paper Session *12:25 to 1:55 pm*

Hilton San Francisco Union Square: Continental 2

Participants:

Effects of Motivation and Self-efficacy on Scientific Literacy Mediated by Science-related Attitude *Sun Geun Baek, Seoul National University; Hye Ji Kil, Chungbuk National University; Anna Shin, Seoul National University*

The effects of achievement motivation and science self-efficacy on scientific literacy meditated by science-related attitude were investigated comparatively with the PISA 2015 data among the USA, South Korea, and Finland. As results, this study's structural equation model fitted satisfactory and there were statistical differences in structural coefficients among these countries.

Evaluating Item Fit Statistic Thresholds for Cross-Country Comparability Analysis in PISA *Seang-Hwane Joo, Educational Testing Service; Lale Khorramdel, National Board of Medical Examiners; Kentaro Yamamoto, Educational Testing Service; Hyo Jeong Shin, educational testing service; Frederic Robin, Educational Testing Service*

The impact of various IRT-based item fit thresholds on the scale and score estimation is examined in PISA 2015. The goal is to account for DIF while establishing measurement invariance at the same time. We found that a threshold of RMSD equals to .10 provides the best results.

Mathematics Performance is Measurement Invariant in PISA 2015 Across the ICT-Participating Countries *Bryce Odell, University of Alberta; Maria Cutumisu, University of Alberta*

This study investigated the measurement invariance of mathematics performance across the 48 countries that participated in the PISA 2015 ICT Familiarity Questionnaire. The findings revealed that mathematics was measurement invariant over the ten plausible MATH values for multiple countries with different cultures. Theoretical and practical implications are discussed.

Trends in student motivation profiles in TIMSS Mathematics across 20 years *Michalis P. Michaelides, University of Cyprus - Dept. of Psychology; Gavin Thomas Lumsden Brown, The University of Auckland; Hanna Eklöf, Umeå University; Elena Papanastasiou, University of Nicosia; Militsa Ivanova, University of Cyprus; Anastasios Markitsis, Independent researcher*

Cluster analysis on items measuring mathematics motivation in TIMSS revealed consistent and inconsistent student profiles. Profiles differed systematically in their relationship with achievement and demographics. Similar patterns emerged across 12 jurisdictions, and in the 1995, 2007 and 2015 administrations despite diverse cultural contexts, and different items measuring motivation across time.

Uncertainty of Item Location for Multistage Adaptive Testing in International Large-Scale Assessments *Yuan-Ling Liaw, UiO CEMO*

One main challenge in ILSA is adequately measuring dozens of highly heterogeneous populations. Multistage adaptive testing offers the possibility of targeting the test to examinees' proficiency levels. With the limited knowledge of item location, probabilistic misrouting offers advantages over merit routing for controlling bias in item parameters and group-level proficiency.

Chair:

Kari Hodge, NACE International Institute

073. Missing Data

Paper Session *12:25 to 1:55 pm*

Hilton San Francisco Union Square: Continental 3

Participants:

Addressing Missing Data in College Surveys: A Person Parameter Recovery Investigation *Shimon Sarraf, Indiana University Center for Postsecondary Research; Dubravka Svetina Valdivia, Indiana University*

We investigate person parameter recovery with the graded response model by applying missing data handling methods to empirical higher education survey data. Specifically, we ask how well traditional and modern missing data handling methods can recover person parameters under various missing proportions and mechanisms (MCAR, MAR, NMAR).

Handling Missingness in Data from Computer-Adaptive Test with the Continuous Response Model *Ramsey Lee Cardwell, University of North Carolina at Greensboro; Geoffrey T LaFlair, Duolingo*

Samejima's (1973) Continuous Response Model estimates parameters for continuous-scale items. A simulation study investigated parameter recovery of the CRM in the presence of missing data. Results indicate that difficulty parameters can be well-recovered even with high missingness when multiple imputation is utilized. Application to real data, however, illustrates obstacles.

Using Multiple Imputations to Handle Non-Random Missing in Multidimensional Multistage Testing *Sinan Yavuz, Mr.; Xiaying (James) Zheng, American Institutes for Research; Young Yee Kim, American Institutes for Research*

In MSTs, when the routing decision is based on the combined ability estimates from multiple subscales, the missingness in adaptive stage is not at random, thus item parameter estimates from multiple unidimensional IRT will be biased. This research proposes using multiple-imputations approach to address this problem via simulations.

Using Response Time Modeling to Detect Speeded Examinees with Missing Responses *Xiaying (James) Zheng, American Institutes for Research; Tong Wu, Purdue University; Young Yee Kim, American Institutes for Research; Fusun Sahin, American Institutes for Research*

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

Chair:

Brian F. Patterson, Curriculum Associates

Discussant:

Joshua Goodman, NCCPA

074. Professional Credentialing Programs

Paper Session *12:25 to 1:55 pm*

Hilton San Francisco Union Square: Franciscan Ballroom A

Participants:

Evaluating Use of an Online Open-Book Resource in a High-Stakes Credentialing Exam *Aaron Myers, University of Arkansas; Bozhidar M. Bashkov, American Board of Internal Medicine*

Despite recent attention to open-book examinations, testing practitioners know little about the conditions in which examinees use open-book resources during high-stakes exams. Examination log file data were investigated to determine examinee and item characteristics associated with use of an open-book resource. Implications for open-book examination development and administration are discussed.

Implementing a New Exam User Interface - Does It Really Make a Difference? *Cecilia Alves, Medical Council of Canada; Andre De Champlain, Medical Council of Canada; Nicole Robert, Medical Council of Canada; Becca Carroll, Medical Council of Canada; Allison Burnett, Medical Council of Canada*

The purpose of this study is to assess the level of comparability of test results across two different User Interfaces (UI) used on a computer-based medical exam. This study explores whether these UI differences significantly impacted the examinee's test-taking experience and item statistics.

Investigating the Use of Polytomous based Testlet in Mixed Format CAT *Ye Ma, University of Iowa; Doyoung Kim, NCSBN; William Muntean, Pearson; Hong Qian, National Council of State Boards of Nursing; Wei Xu, National Council of State Boards of Nursing*

In order to measure higher-order constructs, polytomous item testlet (PIT) is currently being developed. The goal is to investigate the impact of using PIT in CAT from a test design perspective. This study aims to provide insights to practitioners on how to administer PITs in mixed format CAT.

Modeling Item Revisit Behavior: The Hierarchical Speed Accuracy Revisits Model *Ummugul Bezirhan, Teachers College Columbia University; Matthias von Davier, National Board of Medical Examiners; Irina Grabovsky, National Board of Medical Examiners*

This study proposes an extension of the hierarchical speed-accuracy framework for joint modeling of responses, response times and item revisits. The application of the proposed model is demonstrated with a large-scale medical licensure examination. Results suggested that the proposed framework helps capturing examinees' test-taking strategy differences.

Modeling the Conditions Influence Response Time and Performance *Yang Shi, University of California, Berkeley; Richard Feinberg, National Board of Medical Examiners*

This study proposes an extension of the traditional hierarchical speed-accuracy model to check the assumption of conditional independence. Results from a large-scale medical licensure examination demonstrated negative dependency between response time and accuracy, which is a function of item difficulty, examinee ability, and the cognitive processing of time management.

Chair:

Michael Peabody, American Board of Family Medicine

Discussant:

Susan Davis-Becker, ACS Ventures, LLC

075. Non-Cognitive Measures

Paper Session *12:25 to 1:55 pm*

Hilton San Francisco Union Square: Franciscan Ballroom C

Participants:

A Hierarchical Model to Measure Diagnosing Ability and Efficiency in Clinical Diagnosing Reasoning Process *Feiming Li, Zhejiang Normal University; Frank Papa, University of North Texas-Health Science Center; Mengqiu Zhou, Zhejiang Normal University*

Computer-based environment enables assessments of clinical diagnostic reasoning and data gathering skills in more interactive, nonintrusive and less expensive way. This study proposed a hierarchical model to measure both diagnosing accuracy and efficiency underlying the process action data, and investigate the most effective instructions on both dimensions.

Collaboration, Communication, and Content Knowledge: Exploring Performance Predictors in a Simulation-based Assessment *Jessica Andrews Todd, Educational Testing Service; Stephanie Peters, ETS; Carol M Forsyth, Educational Testing Service; Jonathan Steinberg, Educational Testing Service*

In this study, we sought to develop a measure specific to collaborative problem solving (CPS) and explore how CPS competency, communication competence, and content understanding contribute to students' success in a collaborative simulation-based assessment. Results showed content knowledge, responsiveness, and social aspects of CPS positively predicted task success.

Instructional Sensitivity of Non-Cognitive Outcome Measures *Alexander Naumann, DIPF | Leibniz Institute for Research and Information in Education; Burkhard Gniewosz, University of Salzburg; Jan Hochweber, University of Teacher Education, St. Gallen, Switzerland; Johannes Hartig, DIPF | Leibniz Institute for Research and Information in Education, Frankfurt, Germany*

Today, non-cognitive outcomes like motivation receive increasing attention when evaluating teaching effectiveness. However, the instructional sensitivity of such measures is seldomly investigated, leaving the question open whether such measures are capable of capturing effects of instruction. Thus, we aim at extending the concept of instructional sensitivity to non-cognitive outcome measures.

A Computational Psychometric Approach to Measuring Creativity *Denis Dumas, University of Denver; Peter Organisciak, University of Denver; Michael Doherty, Consortium for Interdisciplinary Research on Creativity and Learning*

The reliable and valid measurement of creative thinking has been an elusive goal of psychometricians for decades. This paper details the application of semantic-network algorithms to the quantification of human creativity and compares the psychometric properties of scores derived from various text-mining models to those from human raters.

Chair:

Hao Song, National Board of Certification and Recertification for Nurse Anesthetists

Discussant:

Patrick Kyllonen, Educational Testing Service

076. Electronic Board. Saturday 2:15

Electronic Board Session *2:15 to 3:45 pm*

Hilton San Francisco Union Square: Salon A

Participants:

A Differential Item Functioning Approach to Detect Potentially Compromised Items *Aijun Wang, fsbpt; Yu Zhang, fsbpt; Lorin Mueller, fsbpt*

This study applies the differential item functioning analysis and the person fit statistic to identify potentially compromised items. The proposed approach is applied to both simulated and real data. Results are compared with the results by O'Leary & Smith (2016) method.

Advancing Human Assessment: A Synthesis of the Methodological, Psychological, and Policy Contributions of ETS *Randy E Bennett, ETS*

This presentation synthesizes the many diverse technical, substantive, and policy-related contributions to education and psychology that have occurred over the more than 70-year history of Educational Testing Service (ETS). The talk focuses upon the most consequential of those contributions and their impact on theory and practice.

Correcting the Effects of Missing at Random Data on Polychoric Correlations *Teresa Ober, University of Notre Dame; Alex Brodersen, University of Notre Dame; Ying Cheng, University of Notre Dame*

This study outlines the results of a simulation and a real example to illustrate the issue and proposed methodology for appropriately handling missing data in a polychoric correlation. Preliminary findings suggest imputation of missing at random data may yield accurate results in calculating polychoric correlation coefficients under certain conditions.

EIEvent: Transforming process data into observable outcomes *Russell Almond, Florida State University*

Many modes of assessment, including games, simulation, and video recording, capture event logs. A pattern of events may provide evidence of a target construct. To capture these patterns, the EIEvent system uses analyst-written evidence rules to transform the events into observable variables that can be input into other psychometric models.

Empirical Evidence for Reporting Mathematics Assessment Results Using Cognitive Skill Frameworks *Robert D Ankenmann, University of Iowa; Catherine Welch, University of Iowa; Stephen Dunbar, University of Iowa*

The mandate for statewide standards-based assessment of academic outcomes is accompanied by a desire for summative assessments that provide actionable information to improve instruction and student learning. The purpose of this study is to empirically compare the utility and validity of various cognitive skill frameworks for reporting mathematics assessment results.

Empirical Standard Setting for SEL Measures via the Proposed Interpretive Argument *Michael Clifford Rodriguez, University of Minnesota; Michael Dosedel, University of Minnesota*

By grounding the standard setting approach for measures of social and emotional learning in an interpretation and use argument, we provide a strong basis for supporting the proposed interpretive argument and accomplish validation. We compare a logical-reasoning method of standard setting with ROC analysis and contrasting-groups logistic regression methods.

Evaluating an Automated Scoring System Using Limits of Agreement *Edmund Jones, Cambridge Assessment; Jing Xu, Cambridge Assessment*

Automated scoring of constructed responses is commonly evaluated by passing responses to both the automated system and human raters, and seeing how closely the two agree. We introduce a method from medical science for measuring the degree of agreement, and argue that it is more appropriate than some current methods.

How Quantitative Research and Measurement Departments Address Issues of Racism and Methodology *Katherine Ann Reynolds, Boston College; Michael Russell, Boston College*

This research presents findings from a survey of research and measurement programs regarding how they address issues of racism and methodology in their training of graduate students. Recommendations for how department can engage students with these issues are also provided.

Improving the Efficiency of Virtual Standard Setting Studies *Lucia Liu, Ascend Learning; Matthew Scaruto, Ascend Learning; Kevin Loughlin, Ascend Learning*

This study evaluates critical factors in improving the efficiency of a standard setting using both quantitative and qualitative data collected from a virtual standard setting of an admission test. Retrospective analysis concerning panelist background, panelist ratings, meeting logistics, and panel evaluation provide insight on recruitment, facilitating, and validation.

Investigating Item Position and Response Time Effects with Missing Data *Nayeon Yoo, Teachers College, Columbia University; Ummugul Bezirhan, Teachers College Columbia University; Young-Sun Lee, Teachers College, Columbia University*

The purpose of the study was to examine item position and response time effects when different types of missing data are present. Real-world data analyses were conducted using PISA 2015 Mathematics data, and simulation studies were conducted to incorporate the missing data mechanism into the model.

Language Matters: Stakeholder Perceptions of Achievement Labels from Tests *Francis O'Donnell, University of Massachusetts Amherst; Stephen Sireci, University of Massachusetts Amherst*

Although students, parents, and teachers see achievement labels whenever statewide assessment results are reported, research on how they are perceived is lacking. Based on the belief that better labels can play a role in making measurement matter, we investigated stakeholders' labeling preferences and approaches to discussing results. Recommendations are provided.

Multiple Imputation Approach to Missingness at the Second Level: A Comparison Study *Yifan Bai, American Institutes for Research; Sinan Yavuz, Mr.; Xiaying (James) Zheng, American Institutes for Research; Markus Broer, American Institutes for Research*

NAEP's large missing on background variables gets complicated with its nested structure. This study compares two multiple imputation approaches to address school level missingness: chained equations (MICE) using R and Blimp packages. The results provide empirical suggestions for researchers dealing with multilevel missing data in large-scale assessments like NAEP.

Refined Learning Tracking with A Longitudinal Probabilistic Cognitive Diagnosis Model *Peida Zhan, Zhejiang Normal University; Keren He, Zhejiang Normal University*

To provide a more refined learning diagnosis and tracking, this article proposed a longitudinal probabilistic cognitive diagnosis model (CDM), in which probabilistic attributes are involved instead of binary attributes. The individual differences in attribute mastery and growth rates can be quantified more refined than conventional longitudinal CDMs.

Self-normalization Score-based Tests for Detecting Changing Parameters in Multi-way Clustered Data *Ting Wang, The American Board of Anesthesiology*

It is often difficult to detect changing parameters in multilevel models, manifest as interaction and/or heterogeneity. It gets further complicated if the clusters are defined by more than one grouping factor, such as school and neighborhood. We propose to utilize self-normalization score-based tests to identify the changing parameters in a unified frame.

Tackling the Measurement Specialist Shortage: Empowering Educational Practitioners with Operational Measurement Methodologies *Qinjun Wang, TAL Education Group; Liwei Wei, TAL Education Group; Yuan Zhao D'Antilio, TAL Education Group; Xiujun Yang, TAL Education Group; Chengfeng Wu, TAL Education Group; Angela Bao, TAL Education Group*

This study focuses on the implementation of a standard setting workshop designed to train educational practitioners to perform operational measurement tasks. In addition to the positive outcome found from the workshop, a larger scale survey shows a strong relationship between teacher's attitude toward measurement practices and their general measurement literacy.

The Robustness of the Optimal Item Pool *Tianshu Pan, Pearson*

Reckase (2010) proposed a procedure to create a r-optimal item pool for the fixed-length computerized adaptive test under the normality assumption. This study tries to explore whether the r-optimal item pool works and how the b-parameter is distributed when the ability is non-normally distributed and different stopping rules are used.

The validity of new MCAT scores in predicting USMLE Step 1 scores *Kun Yuan, As sociation of American Medical Colleges; Cynthia Searcy, Association of American Medical Colleges; Andrea Carpentieri, Association of American Medical Colleges*

Total scores from the new MCAT exam predict Step 1 total scores well. MCAT scores and undergraduate grade point averages provide stronger prediction of USMLE Step 1 scores than either predictor alone. MCAT scores also provide comparable prediction of Step 1 scores for students from different backgrounds.

Using Classification Tree Approach for Sub-Score Reporting *Ji Zeng, Michigan Department of Education; Joseph A. Martineau, Center for Assessment*

Using state level assessment data, various classification techniques were compared against each other. Among these, classification tree approach was found to have the best predictive power of the overall score performance levels when classifying students into different groups at domain score level.

Using Graphical Model to Jointly Model Process data and Response Data *Jie Gao, Educational Testing Service; Matthew Johnson, Educational Testing Service*

This paper tries to jointly model the process data and response data in writing assessment, using data from a middle school writing assessment. Using essay length as a sample process feature, the model will evaluate the impact of the process feature on the effect of writing ability on item scores.

Why is This so Hard? Relationships Between Item Features and Angoff Ratings *Kelly Foelber, ABIM; Jerome Clauser, ABIM*

The purpose of this experimental study was to evaluate an intervention aimed at improving standard setting raters' ability to perform the Angoff method. The intervention sought to help raters estimate item difficulty by having them rate the following four features of each item: complexity, familiarity, clarity, and response options.

Ability Estimation in the Presence of Omitted Data in Large-Scale Assessment *R De Ayala, University of Nebraska; Sonia Suarez, University of Nebraska*

Educational measurement typically involves utilizing data that contain missing data such as omitted responses. In this study, data from a large-scale assessment are used to examine the impact on ability estimation of different approaches for handling omitted responses.

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

Coordinated Session – Panel Discussion *2:15 to 3:45 pm*

Hilton San Francisco Union Square: Continental 4

This is one of the two parallel topic sessions at the ATP Innovations in Testing Conference (San Diego, 2020), and at the National Council on Measurement in Education Conference (San Francisco, 2020). The goal of these sessions is to enhance awareness of business vs research needs in the assessment and learning areas. For those working in the assessment industry, there are many competing demands that require constant attention. In many scenarios, these include demands on the time required to complete a project, the financial requirements of the project, and the need to develop or maintain assessments that are consistent with professional standards. During this session, a panel of seasoned measurement and educational technology professionals will discuss scenarios that require excellent judgment and experience to determine how best to meet these competing demands. Come and join us for this session and jump in and share your experiences attempting to juggle requirements to meet professional standards within the practical realities of the world.

Presenters:

Wayne J. Camara, ACT
Susan Davis-Becker, ACS Ventures, LLC
William G. Harris, Association of Test Publishers
John Weiner, PSI Services LLC

Chair:

Jerry Gorham, Ascend Learning

078. **Invited Session The Value of and Values in Educational Assessment**

Coordinated Session *2:15 to 3:45 pm*

Hilton San Francisco Union Square: Imperial Ballroom A

Values underlie all aspects of the establishment, development, and evaluation of educational testing programs. To address current criticisms of educational tests; such as narrowing the curriculum, widening achievement gaps, focusing on unimportant skills, and creating anxiety in children; in this symposium we will discuss the values inherent in current educational testing programs, and how we can reexamine these values to promote more equitable outcomes for students. The symposium will consist of two parts. The first part will comprise three 10-minute presentations by speakers invited to present on current values in educational assessment policy, educational test development, and educational test evaluation. The second part will be a blue-ribbon panel discussion of those topics focused on improving educational equity. Specifically, the panel will respond to the question "How can we re-center our current values in educational assessment to empower traditionally marginalized groups in the educational assessment process?" Dialogue among the panelists, presenters, and the audience will be facilitated. A goal of the symposium is to discover ways in which educational testing can be re-conceptualized to support the learning and progress of all students.

Presenters:

Carl Cohn, Claremont University
Maria Elena Oliveri, Educational Testing Service
Stephen Sireci, University of Massachusetts Amherst

Panelists:

Stafford Hood, University of Illinois
Suzanne Lane, University of Pittsburgh
Darius Prier, Duquesne University
Jennifer Randall, University of Massachusetts Amherst
Cindy Walker, Duquesne University
Amy Stuart Wells, Teachers College

Session Organizer:

Stephen Sireci, University of Massachusetts Amherst

079. Perspectives on Artificial Intelligence in Measurement

Coordinated Session *2:15 to 3:45 pm*

Hilton San Francisco Union Square: Imperial Ballroom B

How should specialists in the educational measurement community adapt to ongoing changes and trends from the area of artificial intelligence (AI)? In this session, we have gathered four speakers from across the educational measurement and AI spectrum to spark a discussion in addressing this question, providing actionable insights to help NCME members reason about how AI might impact their work, careers, and society, measurement, and assessment at large. Each speaker offers a unique perspective on the possibilities of AI and educational measurement, representing a variety of backgrounds that include the testing industry, academia, and the well-funded "pure AI" research industry. All speakers have deep, practical experience in solving current AI problems with their own community's best practices. The format of the session will allow each speaker time to describe how they arrived in the AI field and to elucidate their views on how the field of educational measurement could be going given their knowledge of the AI space.

Participants:

Testing Industry Perspective: Intersection of AI, Measurement, and Applications *Steven Tang, eMetric*

Academia Perspective: Blending AI and Psychometric Principles *Zachary A Pardos, UC Berkeley*

Pure AI Perspective: AI in Specialized Industries *Susan Zhang, OpenAI*

Career Perspective: Transitioning from Measurement to Data Science *SeungYeon Lee, Chan Zuckerberg Institute*

Session Organizer:

Steven Tang, eMetric

Discussant:

Andre Alexander Rupp, Educational Testing Service (ETS)

080. CATs, BATs, and RATs—The Value of CAT for Educational Assessment

Coordinated Session – Panel Discussion *2:15 to 3:45 pm*

Hilton San Francisco Union Square: Yosemite A

Computerized Adaptive Testing (CAT) turns 50 years old in 2020 which may be a shock to many in educational assessment who are still struggling to implement CAT in a way that fully realizes its promised advantages in terms of improved efficiency in testing. Licensure and certification assessment have been leveraging CAT successfully for years. While there have been recent several recent examples of CAT implementations in K-12 summative assessment (such as the Smarter-Balanced Assessment Consortium and Virginia's Standards of Learning assessment), CAT has been relatively slow to catch on in K-12 educational assessment. This is due, in part, to technology limitations and differences between delivering tests to test centers and delivering tests to students in classrooms. However, technology is not the only consideration influencing the effective use of CAT in K-12 assessment. Frequently, constraints are placed on K-12 assessment programs in terms of educational policies, content standards coverage, and comparability that limit the degree to which CAT can deliver assessment efficiently and effectively. This results in assessment programs which are sometimes referred to as "BAT"s (Barely Adaptive Tests) and "RAT"s (Rarely Adaptive Tests). This panel will discuss the challenges associated with CAT in K-12 assessment and forecast its future utility.

Presenters:

Michelle Barrett, ACT, Inc.
Richard Luecht, University of North Carolina
Laurie Laughlin Davis, Curriculum Associates
Walter (Denny) Way, College Board

Session Organizer:

Laurie Laughlin Davis, Curriculum Associates

Chair:

Michael Edwards, Arizona State University

Discussant:

David Thissen, University of North Carolina at Chapel Hill

081. Diagnostic Assessments: Moving from Theory to Practice

Coordinated Session *2:15 to 3:45 pm*

Hilton San Francisco Union Square: Yosemite B

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

Participants:

Designing a Diagnostic Assessment *Leanne R. Ketterlin-Geller, Southern Methodist University*

Weighing Parsimony and Flexibility in Diagnostic Classification Model Selection *Meghan L Fager, National University; Matthew Madison, Clemson University*

Communicating Results of Diagnostic Assessments *Laine Bradshaw, University of Georgia*

Technical Evidence for Diagnostic Assessments *William Thompson, University of Kansas; Amy K Clark, University of Kansas; Brooke Nash, University of Kansas*

Session Organizer:

William Thompson, University of Kansas

Chair:

William Thompson, University of Kansas

Discussant:

Robert Henson, University of North Carolina at Greensboro

082. Identifying Responses that are Unsuitable for Evaluation by Automated Scoring Systems

Coordinated Session *2:15 to 3:45 pm*

Hilton San Francisco Union Square: Yosemite C

In assessment programs, automated scoring systems are frequently used to efficiently score essays and constructed responses. While much research has focused on the overall accuracy, a critical issue in practice using these systems is identifying responses that cannot (or should not) be scored by an automated system. To ensure the validity of automated scoring solutions and the assessments they support, it is vital to know the limits of the automated scoring models and responsibly escalate responses for additional review by human raters. The presentations in this session focus on methods to identify such responses. The first presentation evaluates a technique for identifying responses that contain disturbing content. The second presentation presents a method for routing responses to expert raters based on their suitability for automated scoring using a hypothesis test framework. The third presentation presents an alternate method for escalating responses for human review by using a confidence index that is converted to a percentile based upon a larger second sample. The final presentation discusses how the continuous property of automated score data can be optimized to identify where a scoring engine has low confidence within each score point to escalate responses to humans.

Participants:

MTLHealth: A Disturbing Content Identification System *Erin Yao, ACT; Joseph Valencia, Oregon State University*

A Hypothesis Testing Approach to Optimal Response Routing *Corey Palermo, Measurement Incorporated*

Characteristics of Responses Routed for Human Scoring *Susan Lottridge, AIR*

Smart Routing: Utilizing Within-Score Point Agreement to Escalate Low Confidence Scores to Humans *Sarah Quesen, Pearson; Karen Lochbaum, Pearson*

Session Organizer:

Erin Yao, ACT

Discussant:

Mark David Shermis, American University of Bahrain

083. Methods of Monitoring Human Raters

Coordinated Session *2:15 to 3:45 pm*

Hilton San Francisco Union Square: Continental 1

This session of four papers focuses on an international body of applied research that seeks to better understand and ultimately improve the ways that we monitor and provide feedback to human raters. The first paper, written by Bimpeh, presents an innovative way to monitor raters in an online scoring environment. The second paper, written by Leusner and Wendler, focuses on raters' attitudes about various rater monitoring practices used in operational scoring. The third paper (by Ricker-Pedley) examines the level of consistency of rater performance across multiple types of rater monitoring practices. The fourth paper (by Benton) compares the ratings of senior and junior raters to determine best practices regarding the combining of multiple ratings for the sake of reporting. Rosemary Reshetar of College Board will serve as the Discussant for the session.

Participants:

Identifying errant raters: quality assurance system in human rating *Yaw Bimpeh, AQA*

The Perceived Effectiveness of Rater Feedback *Dawn Leusner, ETS; Cathy Wendler, ETS*

Relationship among performance measures for human constructed response raters *Kathryn Ricker-Pedley, ETS*

Which is better, one experienced assessor or many inexperienced assessors *Tom Benton, Cambridge Assessment*

Session Organizer:

Edward W Wolfe, ETS

Discussant:

Rosemary Reshetar, The College Board

084. Aberrant Testing Behaviors

Paper Session *2:15 to 3:45 pm*

Hilton San Francisco Union Square: Continental 2

Participants:

A Bayesian Approach to Detect Test Fraud *Sandip Sinharay, ETS; Matthew Johnson, Educational Testing Service*

A Bayesian approach was suggested for detecting test fraud. The approach involves the computation of the posterior probability of a better performance on one subtest compared to another. The new approach leads to fewer false alarms and more true positives compared to existing approaches in a simulation study.

Clique-based Detection of Examinees with Preknowledge on Real, Marked Data *Dmitry Belov, LSAC; Sarah L Toton, Caveon*

An experiment was conducted to embed item preknowledge into a group of test takers, where around 50% of them had access to a subset of items before the test. The resulting real dataset was used to study clique-based detectors of item preknowledge employing responses and response times.

Detecting Aberrant Response Behavior in CBTs: A Lognormal RT Testlet Model Approach *Wei Xu, National Council of State Boards of Nursing; William Muntean, Pearson VUE; Doyoung Kim, NCSBN*

Aberrant test-taking behavior can significantly diminish validity of test results. In this study, the performance of the lognormal response time (RT) testlet model is compared with that of the lognormal RT model for detecting aberrant response behavior when local RT dependence is presented.

Influence of Leak Range for a Sequential Compromised Item Detection Procedure *Canxi Cao, Beijing Normal University, PRC.; Jinming Zhang, University of Illinois at Urbana-Champaign*

When examinees have pre-knowledge of the items, their test performance is a no longer valid reflection of examinee knowledge. This study is exploring whether a sequential detection procedure could successfully identify a compromised item under different leak range (the ratio of the compromised item in the pool) conditions in CAT.

Chair:

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

Discussant:

Bradley Thiessen, New College of Florida

085. Response Time

Paper Session *2:15 to 3:45 pm*

Hilton San Francisco Union Square: Continental 3

Participants:

Alternative Approaches to Modeling the Relationship Between Response Time and Accuracy *Haiqin Chen, American Dental Association; Paul De Boeck, Ohio State University; Chien-Lin Yang, American Dental Association; David M Waldschmidt, American Dental Association*

This study investigates the relationship between response time and accuracy using linear and generalized linear mixed modeling approaches. Results reveal that the relationship between response time and accuracy is dependent on item difficulty but independent of the choice of response time transformation or combinations of link functions and distributional assumptions.

Dirichlet Process Mixture IRT Model with Response Time for Aberrant Response Behavior *Juyeon Lee, The University of Georgia; Allan Cohen, University of Georgia*

We explore the usefulness of a Dirichlet process (DP) mixture in polytomous IRT to address aberrant response behavior in questionnaire data. Item response time will be incorporated into the model to assist in interpretation of latent classes.

Exploring the Relation between Item Response and Response Time of the TORRjr Scale *Hongyang Zhao, University of Maryland; Yang Liu, University of Maryland; Patricia Alexander, University of Maryland*

The study explores the factor structure of the item-level response times for the Test of Relational Reasoning-Junior. A joint factor model for item responses and response times is also fitted to reveal the relation between children's latent speed in completing the test and their latent relational reasoning ability.

Impact of Translation on Item Responses and Item Response Time *Hyeon-Joo Oh, Educational Testing Service; Hongwen Guo, Educational Testing Service; Lida Chen, University of Iowa*

In this study, we investigate how the different language versions (i.e., English and Spanish) of the same test affected item function and item response time using differential item functioning (DIF). Also, we conduct DIF analysis with weighted sum scores, as a matching criterion, to evaluate the impact of translation.

Modification of the lognormal response time model to include position effects *Yang Du, University of Illinois, Urbana-Champaign; Xiao Luo, Measured Progress; Xi Wang, Measured Progress; Louis Roussos, Measured Progress*

While the lognormal response time model (van der Linden, 2006) has gained popularity, its constant speed assumption may be problematic in certain test settings. This study investigates the situations where the constant speed assumption is violated and develops a model for the effect of item position on students' speed in low-stakes tests.

Chair:

Yu Fang, Law School Admission Council

Discussant:

Kirk Alan Becker, Pearson

086. Differential Item Functioning #1

Paper Session *2:15 to 3:45 pm*

Hilton San Francisco Union Square: Franciscan Ballroom A

Participants:

An Evaluation of Differential Item Functioning Tests for Multi-Stage Testing *Rudolf Debelak, University of Zurich; Sebastian Appelbaum, Universität Witten/Herdecke; Dries Debeer, University of Zurich; Martin Tomasik, University of Zurich*

This study evaluates three statistical tests for the detection of differential item functioning in the context of multistage tests by the means of a simulation study. A new test based on permutation testing shows promising results when compared with MSTSIB and logistic regression.

An Examination of NCDIF Index for Identifying Anchor Item Parameter Drift *Juan Chen, National Conference of Bar Examiners; Won-Chan Lee, University of Iowa; Mark Connally, National Conference of Bar Examiners; Mark A Albanese, National Conference of Bar Examiners*

This simulation investigated the performance of NCDIF in identifying anchor item parameter drift when sample size, anchor length, percentage of IPD items, IPR procedures, weights, and examinee distributions were manipulated. Preliminary results indicate that NCDIF performed better for shorter forms and different IPR procedures affected the performance for unequal-sample conditions.

Exploring non-traditional latent DIF variables for external characteristics using mixed Rasch *Rongxiu Wu, University of Kentucky; Lijun Shen, University of Kentucky; Chen Qiu, University of Kentucky; Michael Peabody, American Board of Family Medicine*

Non-traditional, latent variables can be more appropriate for identifying the source of DIF than traditional, manifest variables. This study determined the number of latent classes and constructed new, non-traditional latent variables based on patient and practice demographics. The process for constructing and evaluating multiple potential latent variables is described.

Investigating Consistency in Identifying Items with Differential Item Functioning *Xiuyuan Zhang, Psychometrician; Anita Rawls, Thomas Jefferson University*

The present study investigates methods to improve the consistency in identifying items presenting severe Differential Item Functioning (DIF) during an operational administration using semi-simulated data. The findings suggest plausible approaches to more accurately and efficiently estimate an items' DIF status at an earlier stage of the item life cycle.

Testing Differential Item Functioning without Anchor Items *Weimeng Wang, University of Maryland, College Park; Yang Liu, University of Maryland; Hongyun Liu, Beijing Normal University; Ke-Hai Yuan, University of Notre Dame*

Tests for differential item functioning (DIF) often rely on anchor items that set a common metric for the reference and focal groups. We propose a new DIF detection procedure that does not explicitly assume anchor items and evaluate their type I error rate and power via Monte Carlo experiments.

Chair:

Sien Deng, ACT Inc.

Discussant:

Allison Ames, University of Arkansas

087. Assessment and Language

Paper Session *2:15 to 3:45 pm*

Hilton San Francisco Union Square: Franciscan Ballroom C

Participants:

DIF in Native Language and Gender Groups in Language Assessment *Mehmet Kaplan, New Jersey Institute of Technology; Burak Aydin, Recep Tayyip Erdogan University; Ahmet Hattatioglu, Yunus Emre Institute; Tuba Arabaci Atlamaz, Abdullah Gul University; Nursel Tan Elmas, Yunus Emre Institute*

In decades, cultural fairness in language assessment has been a fertile field for researchers. One major concern is whether test items behave differently for different groups. This study aimed to examine and reveal possible sources of bias caused by the same language family and gender groups in language assessment.

Does cognitive fatigue threaten test validity in foreign language assessments? *Christoph Lindner, Faculty of Education, University of Hamburg (UHH), Germany; Jan Retelsdorf, Faculty of Education, University of Hamburg (UHH), Germany*

The present study shows that perceived—and not experimentally manipulated—cognitive fatigue undermines seventh graders' achievement-related outcomes (i.e., low test-taking motivation and performance, more distracting thoughts) during a test of English as a foreign language. We discuss the results in terms of cognitive fatigue as a threat for test validity.

Exploring GRE-TOEFL Score Profiles of International Students Applying to Graduate School *Katrina Crotts Roohr, Educational Testing Service; Margarita Olivera-Aguilar, Educational Testing Service; Jennifer Bochenek, Educational Testing Service; Vinetha Belur, Educational Testing Service*

Although previous studies have examined the factors that international students consider when applying to graduate school, the role of test scores has not been explored. Using finite mixture models, we examined the GRE and TOEFL score profiles of Chinese and Indian students, and their relationship with graduate programs' characteristics.

Simultaneous Linking for Maintaining the Vertical Scale of a Two-Level Test *Jiyun Zu, Educational Testing Service; Lixiong Gu, Educational Testing Service*

We study different linking designs for maintaining vertical scales using real data from a two-level English language test. We hypothesize that it is beneficial to embed both within-level and between-level common items and use simultaneous linking method. Between-level common item parameter estimates, linking coefficients, and new form conversions are compared.

Uses and Adaptations of the Comparative Judgement Methodology for Language Standards Alignment *Sarah Rita Hughes, Pearson; Rose Clesham, Pearson*

This research reports on new uses and adaptations of Comparative Judgement (CJ) as a method of linking an established language framework, PTE Academic's Global Scale of English (GSE), and the newly developed China's Standards of English Language Ability (CSE).

Chair:

Erin Winters, UC Davis

Discussant:

Molly Faulkner-Bond, WestEd

089. **Invited Session Battle of the NCME Past Presidents**

Coordinated Session *4:05 to 6:05 pm*

Hilton San Francisco Union Square: Continental 4

In this session, 19 Past-Presidents of NCME will respond to (mostly serious) questions on topics in educational research and psychometrics. Topics will include Measurement History, Standard Setting, Classical Test Theory, Item Response Theory, and Validity. Questions will be submitted from NCME members in advance. The Past-Presidents will be organized into three teams (west, central, and east) and will compete in a game show-like format to see which region of the country earns the most points. The session promises to be both educational and fun.

Team Central:

Terry Ackerman, University of Iowa
Robert Brennan, Professor Emeritus, University of Iowa, Center for Advanced Studies in Measurement and Assessment
Gregory Cizek, University of North Carolina at Chapel Hill
Michael Kolen, University of Iowa
Mark Reckase, Michigan State University
Barbara Plake, University of Nebraska-Lincoln

Team West

Edward Haertel, Stanford
Lorrie Shepard, University of Colorado
Richard Patz, University of California-Berkeley
Mark Wilson, University of California, Berkeley
Laurie Wise, HumRRO
Rebecca Zwick, Educational Testing Service

Team East

Randy E Bennett, ETS
Wayne J. Camara, ACT
Linda Cook, Educational Testing Service
John Fremer, Caveon
Ron Hambleton, University of Massachusetts - Amherst
Suzanne Lane, University of Pittsburgh
Wim J van der Linden, University of Twente

Judges:

Won-Chan Lee, University of Iowa
Molly Faulkner-Bond, WestEd
Mark Hansen, University of California, Los Angeles

Session Organizer:

Stephen Sireci, University of Massachusetts Amherst

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

Coordinated Session – Panel Discussion *4:05 to 6:05 pm*

Hilton San Francisco Union Square: Imperial Ballroom A

What characterizes quality in tests and testing, and how can that quality be fostered? The federally mandated "Peer Review" of state assessments is a quality control process with specific criteria that will be examined in this session. To provide all attendees essential background, this session will have an initial presentation on Peer Review, which states have undergone for nearly two decades. Then a panel of experts will discuss the nature, impact, strengths and weaknesses, and future of Peer Review. The panel includes persons with expertise in assessment validity, assessment of special populations, a state assessment veteran, a consultant who has helped states comply with Peer Review and other technical criteria for over 20 years, and a U.S. Department of Education officer with responsibilities for Peer Review. The panel format will promote interaction and insight among the panelists, and between the panelists and the audience. Ample time will be allowed for audience participation in the discussion as well.

Presenters:

Brian Gong, Center for Assessment
Don Peasley, U.S. Department of Education
Martha Lynn Thurlow, NCEO/University of Minnesota
Liru Zhang, Delaware Department of Education

Session Organizer:

Liru Zhang, Delaware Department of Education

091. Using Psychometric Approaches to Improve Test Development

Coordinated Session *4:05 to 6:05 pm*

Hilton San Francisco Union Square: Imperial Ballroom B

The quality of assessments is closely tied to the quality of test development processes and procedures. When thoughtful and rigorous test development procedures are lacking, achieving high-quality measurement can be challenging. This session explores several innovative psychometric approaches aimed to improve the entire span of the test development process—from the early stages of practice analysis and blueprinting all the way to examining item quality before and after pretesting. Presenters will share their findings and offer practical advice to attendees in the hope of inspiring improvements to test development and exam programs beyond their own.

Participants:

Developing Useful Test Blueprints: An Evaluation Study *Andrew C. Dwyer, American Board of Pediatrics; Robert Furter, American Board of Pediatrics*

One Blueprint Fits All? Using Person-centered Methodological Approaches in Practice Analysis Studies *Pamela Kaliski, American Board of Internal Medicine; Kelly Foelber, ABIM; Jerome Clauser, ABIM*

The Effect of Item Writing Experience on Item Quality *Bozhidar M. Bashkov, American Board of Internal Medicine; Jerome Clauser, ABIM*

Modeling Text Complexity Judgments and Their Association with Item Difficulty *Brian F. Patterson, Curriculum Associates; Pamela Seastrand, Curriculum Associates*

Using Item Text to Predict Item Survival: Can It Go Beyond Linguistic Characteristics *Victoria Yaneva, National Board of Medical Examiners; Peter Baldwin, National Board of Medical Examiners (NBME); Janet Mee, National Board of Medical Examiners*

Session Organizer:

Bozhidar M. Bashkov, American Board of Internal Medicine

Discussant:

Kurt F. Geisinger, Buros Center for Testing

092. Making Measurement Matter by Unpacking Cognitive Complexity: What is it and Why is it so Hard?

Coordinated Session – Panel Discussion *4:05 to 6:05 pm*

Hilton San Francisco Union Square: Yosemite A

Every item that is used in large-scale k-12 assessment must be coded for depth of knowledge. The DOK is used in form construction to ensure that a breadth of cognitive demands is represented on the assessment. Alignment studies review DOK for items and learning standards to evaluate match. However, there is a call in the field for a different lens through which to view cognitive complexity and to more explicitly incorporate cognitive complexity into assessment design and score interpretation (Ferrara, 2017; Ferrara, Steedle, & Frantz, 2018; Ferrara, et al., 2014; Huff & Plake, 2010; Nichols & Huff, 2017; Perie & Huff, 2016; Schneider, et al., 2013). This body of research reflects a number of persistent challenges with respect to cognitive complexity: conceptualization, measurement, and implementation. The panel discussion will address and discuss these challenges and put forth innovative solutions to each.

Presenters:

Ellen Forte, edCount, LLC
Christina Schneider, NWEA
Steve Ferrara, Measured Progress
James Pellegrino, University of Illinois, Chicago

Session Organizer:

Kristen L Huff, Curriculum Associates

Discussant:

Kristen Huff, Curriculum Associates

093. Innovations in Assessment: An International Perspective

Coordinated Session *4:05 to 6:05 pm*

Hilton San Francisco Union Square: Yosemite B

All over the world, assessments are used to make meaningful decisions on student learning. Depending on the particular situation, these assessments have high impact for individuals or aim to inform policy decisions on group levels. Also, these assessments are grounded in different measurement paradigms that are aligned with cultural perspectives and characteristics of educational systems. All these systems hold their own challenges when it comes to well-known measurement problems like comparability, reliability and validity. This coordinated session aims to facilitate a discussion on innovations in assessment in six different international contexts. It provides new or other perspectives and shows innovative approaches or creative solutions to beforementioned measurement problems. The countries represented in this session are (in alphabetical order): •China: The use of artificial intelligence to analyze data of a music performance test. •France: Analysis of process data from technology enhanced items. •Israel: Research and Practice of Assessment in MOOCs •Japan: CBT-readiness in Japan and the concept of fairness in educational testing and assessment •Sweden: Digitalization of national tests and automated item scoring of constructed responses. •The Netherlands: High stakes adaptive testing and item review solutions.

Participants:

The Use of Artificial Intelligence to Analyze Data of a Music Performance Test. *Tao Xin, Beijing Normal University, PRC.; Jiahui Zhang, the Collaborative Innovation Center of Assessment for Basic Education Quality at Beijing Normal University*

Analysis of Process Data from Technology Enhanced Items. *Thierry Rocher, DEPP, Ministry of education, France*

Research and Practice of Assessment in MOOCs *AVI ALLALOUF, NITE; Giora Alexandron, Weizmann Institute of Science; Anat Ben Simon, NITE – National Institute for Testing and Evaluation; Saar Gershon, Weizmann Institute of Science; Lihi Nahum, Hebrew University of Jerusalem*

CBT-Readiness in Japan and the Concept of Fairness in Educational Testing and Assessment *Louis Liu, Gakken Research Institute for Learning and Education (GRI)*

Digitalization of National Tests and Automated Item Scoring of Constructed Responses. *Anna Lind Pantzare, UMEA University, Sweden*

High Stakes Adaptive Testing and Item Review Solutions. *Hendrik Straat, Cito*

Session Organizer:

Saskia Wools, Cito

Chair:

Saskia Wools, Cito

094. Issues and Considerations for Measuring the Efficacy of Edtech in Higher Education

Coordinated Session *4:05 to 6:05 pm*

Hilton San Francisco Union Square: Yosemite C

Educators, developers, and researchers alike continue to grapple with measuring what works, for whom, and why in educational technology. There are many challenges measurement challenges associated with edtech efficacy including constructing instruments that appropriately capture off-platform data, extracting the right data from platforms given that the tools were not developed to enable measurable outcomes, varied implementations by instructors, and lack of valid and reliable measures. In this session we outline these issues and discuss considerations for sharpening measurement of efficacy through creating appropriate data lakes, constructing valid and reliable measures, and applying the appropriate analyses. Using a next-generation learning system of aligned content, resources, and assessments, this symposium addresses three main areas. First, we discuss how a well-constructed learning model built on proven learning science principles sets the foundation for valid and reliable measurement when in use. Next, we present methods and measures for developing an evolving portfolio of increasingly rigorous evidence to validate the learning model and the impact of the learning tools. Finally, we discuss how learning analytics are leveraged to contribute to the efficacy portfolio. Practical tools that can be implemented by other researchers grappling with edtech efficacy measurement issues will be shared.

Presenters:

Becca Runyon, Macmillan Learning
Jeff Bergin, Macmillan Learning
Marcy Baughman, Macmillan Learning

Session Organizer:

Kara McWilliams, Macmillan Learning

Discussant:

Sara J. Finney, James Madison University

095. Research Blitz - Item Response Theory

Research Blitz Session *4:05 to 6:05 pm*

Hilton San Francisco Union Square: Continental 2

Participants:

A higher-order IRT model with longitudinal data *Tyler Holmes Matta, Pearson; Daniel Furr, Pearson*

This paper presents a higher-order item response models designed to track multidimensional change. By explicitly defining the subdomains and change process in the measurement model makes explicit any assumptions about the domain structure over time, produces improved psychometric properties for all scores, and improves the interpretation of growth.

An IRTree Model for Response Styles and Trait-based Unfolding Responses *Siqi He, University of Illinois Urbana-Champaign; Justin L. Kern, University of Illinois at Urbana-Champaign*

In this study, an extension of IRTree model is proposed under a hierarchical framework to accommodate extreme (ERS) and midpoint response styles (MRS) for unfolding Likert-scale items. The model parameters can be estimated via Bayesian Markov chain Monte Carlo algorithm.

Evaluate Item Response Functions and Test Scores between 3PL and 2PL models *Lin Wang, ETS; Tsung-Han Ho, ETS*

This study investigated whether 2PL and 3PL IRT models would yield comparable item characteristic curves (ICCs) on the same 3PL-based simulated responses; the simulation considered different abilities, test lengths, and sample sizes. The findings showed trivial differences between the 2PL and 3PL results under different simulated conditions.

Impact of Test Lengths on Item Parameter Drift in CAT *CHANGJIANG WANG, Pearson; David Shin, Pearson; Lingyun Gao, Acend Learning*

In response to the trend of reducing testing, the study proposes to investigate the impact of reduced test lengths on the detection of item parameter drift in computer adaptive tests. Preliminary results indicate reduced detectability and increased false negatives. Thus, caution must be used before such educational decisions are made.

Improving Estimation of the Four-Parameter Graded Response Model Using PyMC3 *Christian Meyer, University of Maryland, College Park; Tessa Johnson, University of Maryland, College Park; Yang Liu, University of Maryland; Hong Jiao, University of Maryland*

The four-parameter graded response model (4PGRM), a generalization of Samejima's (1969) "two-parameter" graded response model (2PGRM), fits response patterns for ordinal polytomous items with lower and upper asymptotes of the characteristic curves for the response categories. The current article presents an updated and improved estimation method for the model.

IRT Calibrations of Paired Items on Large-Scale Test *Weiwei Cui, College Board; Liam Duffy, College Board; Daniel Lee, College Board; Amy Hendrickson, College Board*

The response to a paired question is conditioned on the test taker's response to the previous question. The use of paired items violates the local independence assumption which is critical for IRT models. This study focuses on alternative methods to accommodate the local dependence of item pairs.

Item Difficulty Modeling *Emily Bo, NWEA; Sarah Miller, NWEA*

This study proposes to use the random forest approach to set preliminary item difficulties of new item development in MAP Growth. Our preliminary results show a high prediction accuracy that the correlation between the true and the predicted values is 0.93.

Key Content Factors Behind IRT Properties of Verbal Items *Yanyan Fu, GMAC; Kyung Han, GMAC*

The study examined the factors behind the properties of reading passage items that may impact the item parameters. The results suggest that the content category and number of unique words of the passage, and the overlapping nouns between the correct option and incorrect options of an item could influence a-parameters.

Latent Space Item Response Modeling Approach for Binary and Continuous Item Responses *Eric Ming-Yin Ho, UCLA; Minjeong Jeon, UCLA*

We introduce a new latent space modeling approach to item response data that does not need the local independence assumptions for items and respondents. Unobserved dependence among items and respondents are directly modeled as a network. We demonstrate this approach for binary and continuous item responses with empirical illustrations.

Mixture Item Response Theory Model Selection with Hybrid Ant Colony Optimization Algorithm *Zeyuan Jing, University of Florida; Huan Kuang, University of Florida; Walter Leite, University of Florida; A. Corinne Huggins-Manley, University of Florida*

Mixture item response theory model (IRT) has been proposed for handling the impact of latent subpopulation characteristics on the measurement invariance property in psychometrics. In this study, the hybrid ant colony optimization algorithm (hACO) was introduced to select predictors and determine the number of latent subpopulations.

Setting grade standards with IRT: Higher Education Instructor and Student Perspectives *Gavin Thomas Lumsden Brown, The University of Auckland; Paul Denny, The University of Auckland*

SmartStandardSet is a user-friendly application of 2PL IRT score analysis to support course administrators in setting cut-scores for grade boundaries. Students and instructors in beta testing report positive attitudes towards the system and IRT score adjustment but had concerns about ensuring quality in standard-setting decisions.

Testing Latent Variable Distribution Fit in IRT Using Posterior Residuals *Scott Monroe, University of Massachusetts Amherst*

This study proposes a new test statistic for evaluating latent variable distribution fit in IRT, based on posterior residuals (Haberman, Sinharay, & Chon, 2013). A simulation study is used to compare the proposed statistic to the M_2 (Maydeu-Olivares & Joe, 2006) and S-D^2 (Li & Cai, 2017) statistics.

Chair:

Susan Davis-Becker, ACS Ventures, LLC

096. Research Blitz - Validity

Research Blitz Session *4:05 to 6:05 pm*

Hilton San Francisco Union Square: Continental 3

Participants:

A measure of diversity *Kory Vue, University of Minnesota; Tai Do, University of Minnesota; Kyle Nickodem, University of Minnesota; Michael Clifford Rodriguez, University of Minnesota*

Diversity, or the relative heterogeneity of the population (White, 1986), is a very conspicuous topic in schools. Using an appropriate measure to examine diversity would provide evidence to support and evaluate the topic of diversity.

A social-emotional skill assessment for predicting college success and increasing student diversity *Kate Walton, ACT, Inc; Cristina Anguiano-Carrasco, ACT, Inc.; Krista Mattern, ACT, Inc; Jeremy Burrus, ACT, Inc*

This study describes the validation of a social and emotional skills assessment for use in evaluating the college readiness of students. In addition to documenting reliability and validity evidence, we investigated subgroup differences and found negligible differences. Benefits in terms of increased student diversity are discussed.

College Aspirations Prediction using Social Emotional Learning Measures: A Machine Learning Approach *Mireya Smith, University of Minnesota; Alejandra Miranda, University of Minnesota; Ozge Ersan, University of Minnesota; Michael Clifford Rodriguez, University of Minnesota*

Using supervised machine learning model (criterion-related validation), six social emotional learning measures were used to develop an algorithm to effectively predict future college attendance classification. To measure predictive accuracy of the model, testing data results were examined. We recommend using five measures and seven tree splits to predict external variable.

Do different Response Formats Matter in the Measurement of Learning Strategies? *Ana Tupac-Yupanqui, Technical University of Munich, TUM School of Education, Centre for International Student Assessment; Anja Schiepe-Tiska, Technical University of Munich, TUM School of Education, Centre for International Student Assessment; Jörg-Henrik Heine, Technical University of Munich, TUM School of Education, Centre for International Student Assessment; Dave Zes, Long Beach, California; Kristina Reiss, Technical University of Munich, TUM School of Education, Centre for International Student Assessment*

The objective of the present study is to explore the validity of the two response formats, forced-choice versus rating-scale response format, in self-reported questionnaires. This will be examined in learning strategies and their relationship to mathematical competence in adolescent students in the Programme for International Student Assessment (PISA) 2012.

Evaluating Mode Comparability Using Propensity Score Matching: Overcoming Practical Obstacles *Xi Wang, Measured Progress; Louis Roussos, Measured Progress; Fei Chen, University of North Carolina at Chapel Hill*

This study concerns the use of propensity score matching to evaluate and adjust for construct-irrelevant variance introduced by testing mode. We discuss practical challenges from both statistics and policy perspectives and propose new methods to manage them. A demonstration of our methods to a real K-12 testing program is presented.

Examining Evidence for the Validity of PISA 2015 Collaborative Problem-Solving Measure *Sofia Eleftheriadou, University of Manchester*

The validity of the PISA 2015 collaborative problem-solving measure is investigated using a unidimensional Rasch model. Findings indicate that an overall measure of CPS achievement can be supported, but there are also potentially meaningful sub-scales along with the overall measure. Measurement invariance is established for all items across gender.

Examining the Impact of a Consensus Approach to Content Alignment Studies *Sebastian Moncaleano, Boston College; Michael Russell, Boston College*

Although both content alignment and standard-setting procedures rely on content-expert panel judgements, only the latter employs discussion among panel members. This paper examines how a consensus approach to content alignment can enhance the validity of panel judgements. Findings suggest that informed discussion results in more accurate judgements of content alignment.

Exploring 'Ambitious Claims' from a Stakeholders' Perspective when Validating High Stakes Tests *Jerome De Lisle, The University of the West Indies; Tracey M Lucas, The University of the West Indies*

Ambitious test claims require stronger evidence in argument-based validation. Validating ambitious claims may provide a pathway to judging utility and adequacy of 11+ educational policy. Perceptions of ambitious claims for the high stakes 11+ examination were gathered from teachers. They are mostly focused upon construct measurement and unintended negative consequences.

Interval Validation Method Pool Reduction and Domain Balance Using Real Item Pools *William Insko, HMH*

The Interval Validation Method for setting achievement level standards is specifically designed for assessments with large item pools such as computerized adaptive tests. The present study uses real item pools to validate data reduction and content balancing procedures. Recommendations for controlling content coverage for Exemplar items are discussed.

Making Response Processes Matter to Designing Assessments of Student Learning *Weeraphat Suksiri, University of California-Berkeley; Linda Morell, University of California, Berkeley; Mark Wilson, University of California, Berkeley*

Crucial to ensuring the quality of an assessment is verifying that respondents engage in the theoretical processes as intended. This paper discusses two studies in a cognitive domain and a social-emotional domain that illustrate how evidence based on response processes can improve the overall validity investigation.

Using Scale Anchoring to Assist the Content Validity Study of Large-scale Assessments *Michelle Chen, Paragon Testing Enterprises*

Seeking evidence to support content validity is essential in test development and maintenance. By evaluating the adequacy of test content coverage and the alignment of items to the performance standards, this study demonstrates how scale anchoring method could help the design and implementation of content validation studies for large-scale assessments.

Validating Learning Progressions Using Bayesian Inference Networks *Richard D. Schwarz, ETS*

The use of Bayesian Inference Networks for validating learning progressions are demonstrated. This paper encompasses approaches for estimation, model fit, and revisions to a hypothesized learning progression using a classroom-based assessment. The use of content experts in constructing the learning progression and approaches for further model refinement are presented.

Chair:
Andrew Wiley, ACS Ventures, LLC

097. Research Blitz - Test Development Research

Research Blitz Session *4:05 to 6:05 pm*

Hilton San Francisco Union Square: Franciscan Ballroom A

Participants:

An Evaluation of Automatic Item Generation Approach for Generating Parallel Forms *Hongwook Suh, Nebraska Department of Education; Minsung Kim, Buros Center for Testing; Jaehwa Choi, George Washington University; Ji Hoon Ryoo, University of Southern California, Children's Hospital Los Angeles*

A utility of Automatic Item Generation (AIG), parallel or equivalent form generator, is examined on parallel form reliability indices of multiple forms developed within AIG framework. Parameter estimates from an empirical dataset were examined, and a simulation study was conducted with the conditions mimicking the empirical study.

Analyzing Success on a CPS Task Using Response Processes *Yuting Han, Beijing Normal University; Mark Wilson, University of California, Berkeley*

This study provides an analytical framework for analyzing success data from novel scenario-based tasks. The Olive oil task from the ATC21S project is taken as an example to illustrate how to define and score items based on a process diagram. The estimates were obtained using the Rasch model.

Applying Linguistic Frameworks to Reduce Construct Irrelevance on High Stakes Mathematics Items *Kevin Close, Arizona State University; Pamela Paek, ACT Inc.*

Mathematics contents standards require real-world contexts, which threaten to increase conflation between mathematics and linguistic ability on high-stakes assessment items. This study demonstrates how creating linguistic profiles using a qualitative framework and multiple quantitative readability measures, detect and target these conflating features that current item development protocols may overlook.

Change of understanding linear graphs in physics and economics in higher education *Sebastian Brueckner, Johannes Gutenberg-University Mainz; Olga Zlatkin-Troitschanskaia, Johannes Gutenberg-University Mainz*

The (inter-)disciplinary nature of graph understanding and its change over the course of studies have hardly been researched so far. In a pre-post eye-tracking study, the understanding of linear graphs among 34 freshman students of physics and economics from two universities is investigated using isomorphic task pairs for both domains.

Combining think aloud and eye tracking methods for assessment task cognition evaluations *Yu Su, ACT; Carrie A. Morris, University of Iowa; Jay Thomas, ACT, Inc.*

This study is designed to compare data quality from concurrent think aloud (CTA) methods and retrospective think aloud (RTA) methods cued with eye tracking data. Twenty participants will complete cognitively matched forms of an assessment of workplace problem solving with alternating CTA and cued RTA conditions.

Designing and Modeling Test Items for Small Groups *Peter Halpin, UNC*

Research in cooperative learning has addressed how to design tasks that support small group interactions among students. The present research considers how these task designs can be used to "collaborify" conventional test items. Three new item designs are illustrated, pilot data are presented, and an IRT-based modeling framework is proposed.

Evaluating the Impact of a Touchscreen Device in Assessment via Coglab Methods *Kristin Morrison, Curriculum Associates*

A think aloud study was conducted to evaluate and identify potential device effects on student responses when solving assessment items. Students were presented with multiple-choice and writing items on two different devices – computer and iPad. Preliminary results suggest students prefer using the computer and warrant further investigation.

Exploring Online Assessment Personalization via Automated Test Assembly *Gulsah Gurkan, Boston College; Michael Chajewski, Kaplan Test Prep*

ATA has been used for varied of formative goals in online settings. There herein presented study, through an empirical simulation, evaluated two ATA methods for the creation of through-course assessments paired with learners' self-paced QBank content practices. Assessment constraint and item pool usage implications are discussed.

Form Assembly with Standard Setting Item Ratings Masquerading as Empirical Item Difficulties *Corina M Owens, Alpine Testing Solutions; Jill van den Heuvel, Alpine Testing Solutions; Angelica Rankin, Alpine Testing Solutions; Casey Johnson, Alpine Testing Solutions*

Assessment programs must frequently utilize "creative" test development approaches (e.g., they may lack the luxury of conducting beta testing before standard setting). How do items and ultimately forms perform when only standard setting ratings are available? Do ratings sufficiently approximate empirical item difficulties? What can programs learn from one another?

Rasch and Traditional Approaches to Credentialing Examination Job-Task Analysis: A Comparison *Yixi Wang, The Ohio State University; Bridget C. McHugh, The Ohio State University; James Austin, The Ohio State University*

Job verification surveys are widely used in credentialing test design. This study compares Rasch and traditional data analyzing approaches, and compares composite and importance scales in job verification survey. The results indicated strong congruence between two approaches, and reported moderate-strong level of correlations between composite scale and importance scale.

Using Latent Class Analysis for "Pick-N" Items: Creating "Classes" of Learners *Magdalen Beiting-Parrish, CUNY Graduate Center; Sydne T McCluskey, CUNY Graduate Center; Jay Verkuilen, CUNY Graduate Center; Howard Everson, City University of New York; Claire Wladis, Borough of Manhattan Community College*

Multiple Answer Multiple Choice items (MAMC) have been used primarily for high-stakes assessments but are being used in broader circumstances. Unfortunately, they are a complex format for examinees, item writers, and scorers. This proposal uses Latent Class Analysis (LCA) to better understand how examinees respond to these item formats.

Transforming a Region: Designing a Digital Assessment System from Scratch *Bryan R. Drost, NCME*

This exploratory study examined the process by which a a regional consortium of school districts created a digital formative assessment system that is proving to be predictive in relationship to the state assessment system. Analysis of the results showed that systematic professional development related to item creation, gatekeepers on item creation, and blueprint design were key to helping create predictive items. Data also revealed that the

participants were most willing to make instructional adjustments when they understood why and how items aligned to various content standards.

Chair:

Tracy Lynn Gardner, New Meridian Corporation

098. Research Blitz - Test Equating Research

Research Blitz Session *4:05 to 6:05 pm*

Hilton San Francisco Union Square: Franciscan Ballroom C

Participants:

A Bootstrapping Examination of Anchor Size and Equating Variance *Erin Banjanovic, Pearson; Beth Bynum, HumRRO; Marc Kroopnick, Association of American Medical Colleges*

In the world of post-equating, item drift can disrupt the most carefully planned equating designs. The current study aimed to understand how random drift may impact equating through a bootstrapping examination of the equating error associated with three anchor set sizes.

Addressing imprecision and nonoverlap in covariate-adjusted test linking for international contexts *Masha Bertling, Harvard University*

While the matching methodology is becoming widely used in linking literature to compose equivalent groups, the majority of studies are conducted in Western economies with well-validated large-scale assessments. This study compares the performance of matching methods in a novel setting of developing countries and discusses potential drawbacks and benefits under these less than perfect equating conditions

Are Level Information Targets Necessary for Insuring Level Score Precision *MIN WANG, ACT; Meichu Fan, ACT*

To enhance the efficiency and efficacy of test development, using empirical information and simulation approach, this study evaluates the feasibility of relaxing performance-level IRT information targets during assembling parallel forms for an assessment reporting instant performance level scores. The results could provide recommendations for test construction and new form review.

Detection of Outliers in Anchor Items Using Modified Rasch Fit Statistics *Chunyan Liu, National Board of Medical Examiners; Daniel Jurich, National Board of Medical Examiners; Carol Morrison, National Board of Medical Examiners; Irina Grabovsky, National Board of Medical Examiners*

Existence of outlier anchor items could increase equating bias and jeopardize the validity of test scores. This study investigates the performance of three methods to detect anchor outliers in IRT equating. The results indicated that INFIT, OUTFIT, and t-test statistics displayed both high power and low Type I error.

Developing Vertical Scale with Simultaneous Linking Approach *Lixiong Gu, Educational Testing Service*

Vertical scales are typically established using IRT-based approaches, concurrent calibration or Stocking-Lord linking, to place item parameters of all levels onto the same scale. This simulation study evaluates the performance of an alternative simultaneous linking approach, which estimates linking constants simultaneously for separately calibrated test forms, in developing vertical scales.

Effectiveness of Circle-Arc and Rasch Equating for Credentialing Exams with Small Volumes *Amanda Wolkowitz, Alpine Testing Solutions, Inc.; Keith Wright, The Enrollment Management Association*

Previous research has investigated methods for equating at a single pass/fail score using small sample sizes versus equating scores across an entire score scale. This research will focus on the circle-arc and IRT methods, use results from both real and simulated data, and discuss the practical implications of the results.

Equating Oral Reading Fluency Scores: A Model-Based Approach *Akihito Kamata, Southern Methodist University; Yusuf Kara, Southern Methodist University; Cornelis Potgieter, Texas Christian University; Joseph Nese, University of Oregon*

This study demonstrates and evaluates equating procedures for the oral reading fluency (ORF) scores estimated by a latent binomial-lognormal joint model. Preliminary results showed that model-based ORF scores should be preferred over observed words correct per minute measures for passages that have been equated by the common-item non-equivalent group design.

Exploring Adequate Sample Size for Rasch Common Item Equating *Zhen Li, Texas Education Agency; Haiqin Chen, American Dental Association*

This paper investigates the smaller than recommended sample sizes that provide acceptable precision for Rasch equating at the cut points when adding more items around the cut points. The results completed show promise when n is as small as 100 under some study conditions.

Impact of equating with small samples and short tests in formative assessment *Jinah Choi, Edmentum, Inc.; Donna Butterbaugh, Edmentum, Inc.*

Building on previous common item non-equivalent group design equating methodology for adjusting test form difficulty with small sample research, this study examines how characteristics of small-scale formative-assessments (e.g., non-normal distribution of scores, few items per form) influence equating results through simulations and empirical analysis.

SiGNET Circle-Arc Pre-Equating in Small Scale Testing Programs *Alvaro J Arce, Pearson Assessment*

The paper documents the performance of circle-arc methodology for pre-equating SiGNET test forms with multiple conditions of middle-point locations, small sample sizes, and population equating functions. Middle-point location showed the greatest effect on pre-equated test scores when the population equating function was other than mean equating. Small sample sizes worsened the performance of the equating method, and no equating is a better choice than equating.

Supporting small-scale equating: Detecting item parameter drift *Colby Nesbitt, HumRRO; Matthew Swain, HumRRO; Beth Bynum, HumRRO; Marc Kroopnick, Association of American Medical Colleges; Ying Jin, Association of American Medical Colleges; Laurie Wise, HumRRO*

This study presents and evaluates a novel method for identifying item parameter drift in small-scale test administrations. The method compares the observed p-value to the p-value that is expected based on sample ability. Drifting items are identified by percentile thresholds of the full probability distribution of the expected p-value statistic.

Chair:

Samuel Haring, ACT

088. GSIC - Electronic Board. Saturday 4:05

Graduate Student Issues Committee (GSIC)
Graduate Electronic Board Session
4:05 to 6:05 pm
Hilton San Francisco Union Square: Salon A
Participants:

Application of a Bayesian nonparametric IRT Using beta-mixture in addressing atypical ICCs *Juyeon Lee, The University of Georgia; Allan Cohen, University of Georgia*

In this study, we investigate the performance of a nonparametric Bayesian IRT model with a beta-mixture to model atypical ICCs that represent a plateau at the middle of the ability scale. An empirical illustration and simulated data will be provided.

Applying the Item Descriptor Matching Method to Alternate Assessments: Evaluating Panelists' Experiences *Kelley Wheeler, ACS Ventures*

The Item Descriptor Matching method of standard-setting seeks to reduce the cognitive load of panelists. Panelists' understanding of the method is important for the validity of their recommendations. This study evaluates the application of the method to a statewide alternate assessment with recommendations for improving validity evidence of cut scores.

A Study of a Fit Index for Explanatory Item Response Theory Models *Heather Handy, Georgia Institute of Technology*

Applying explanatory item response theory (IRT) models is advantageous when designing and selecting items. A simulation study with varying conditions was conducted to compare an explanatory full information IRT fit index to traditionally used regression-based indices based on limited information for assessing model quality.

Automating Test Administration Decisions in Computerized Formative Assessment *Jinnie Shin, University of Alberta; Fu Chen, University of Alberta; Chang Lu, University of Alberta; Okan Bulut, University of Alberta*

This study introduces a systematic approach to automate test administration decisions for computerized tests. We used time-series prediction and clustering algorithms to demonstrate the feasibility of our approach to automating the test administration process. Our approach achieved high accuracy with interpretable clustering results to maximize the benefits of computerized testing.

Conditional Standard Errors of Measurement for Scale Scores Using MIRT *Hacer Karamese, University of Iowa; Won-Chan Lee, University of Iowa*

In this paper, multidimensional item response theory is considered as a framework for estimating scale score conditional standard errors of measurement (CSEMs), where scale scores are nonlinear transformation of number correct scores. The UIRT and MIRT procedures are applied to real and simulated multidimensional data to estimate CSEMs.

Dimensionality Analyses of a Licensure Test *Tong Wu, Purdue University; Ting Xu, AICPA*

Exploratory and confirmatory factor analyses were conducted to investigate the dimensionality of a high-stake licensure exam that assumes a single latent trait. Preliminary results provide empirical evidence on the robustness of unidimensional IRT models in the presence of multidimensional data.

Evaluating Measurement Bias for Emergent Bilingual Subpopulations *Susan Rowe, University of California, Davis; Megan E Welsh, University of California, Davis*

Typically, IRT and DIF analyses comparing emergent bilinguals (EBs) to English proficient students (EPs) do not account for the heterogeneity of EBs. This paper introduces a method of exploring bias in EBs by conducting IRT and DIF analyses between long-term EBs and EPs and between short-term EBs and EPs.

Examining How to Define Posterior Distributions in IRT Characteristic Curve Linking Methods *Tong Wu, University of North Carolina Charlotte; Stella Yun Kim, UNC Charlotte*

The primary purpose of the study is to examine the effect of seven variations of defining an ability distribution for the IRT characteristic curve linking methods. Simulation study is conducted to compare the relative accuracy of the seven approaches for Haebara and Stocking-Lord methods under a set of simulation conditions.

Factors Affecting the Classification Accuracy in Cognitive Diagnosis Based on BP Neural Network *CHANG NIE, Beijing Normal University; Tao Xin, Beijing Normal University, PRC.*

This research investigates how the five factors affect the classification accuracy in cognitive diagnosis based on BP neural network. The results show the number of attributes and attribute hierarchy have negative impacts, items quality and the test length have significant positive effects, and the sample size does not matter much.

Impact of Choice of Theta Estimator on Student Proficiency Scores *Jennifer Richardson, University of Connecticut*

The new M-MLE estimator was compared to EAP and MAP to assess bias and accuracy in theta recovery under unidimensional Rasch and Rasch testlet models. Preliminary results: testlet based EAP was less biased and more accurate than MAP estimators. MMLE program is still under development. Implications for score interpretation considered.

Item Preknowledge Detection Using Response and Response Time *Onur Demirkaya, University of Illinois at Urbana-Champaign; Jinming Zhang, University of Illinois at Urbana-Champaign*

This study proposes a complimentary analysis by utilizing response time data in addition to response data to build upon Sinharay's signed likelihood ratio statistic (2017) for detection of item preknowledge. Performance of the procedure is measured by the truncated receiver operating characteristic (ROC) curve for both adaptive and nonadaptive tests.

Measurement Invariance for Noncognitive Measures across Unique Populations by Explanatory IRT *Ozge Ersan, University of Minnesota; Michael Clifford Rodriguez, University of Minnesota*

Using a partial-credit explanatory IRT, we examined validation concerns of score interpretation of Family/Community Supports across Somali, Hmong, and American Indian students comparing to white students. We found significant interactions between item thresholds and unique populations implying different perceptions of items and their potential impact on score interpretation.

Modeling Rater Drift: A Covariate-Based Longitudinal Framework Using Latent Class Signal Detection Theory *QIAO LIN, University of*

Illinois at Chicago; Kuan Xing, University of Illinois at Chicago; Yoon Soo Park, University of Illinois at Chicago

This study introduces a covariate-based longitudinal framework using latent class signal detection theory to detect rater drift in constructed response scoring. Real-world data analysis identified changes in rater behavior over time in terms of discrimination ability and severity level. Simulation studies provide inferences on estimation and parameter recovery.

Refining the Q-matrix in Cognitive Diagnostic Assessments Using Wald and Score Tests *ABDULLAH ASILKALKAN, The University of Alabama; Wenchao Ma, The University of Alabama*

Cognitive diagnosis models (CDMs) intend to group individuals into latent classes with distinct attribute profiles to indicate which attribute individuals have possessed and which they have not. CDM analyses could provide diagnostic information about a student's strengths and weaknesses on a set of fine-grained skills to facilitate instruction and learning.

Research on Termination Rules of Multidimensional Computerized Classification Testing *HE REN, Beijing Normal University; Ping Chen, Beijing Normal University*

This study aims to propose new termination rules for multidimensional computerized classification test (MCCT) and explore the possibility of constructing new rules based on machine learning algorithm. The newly proposed rules are thoroughly compared with existing methods through simulations. The results can help improve the efficiency and accuracy of MCCT.

Test Characteristics Influencing the Relative Precision between Summed-score EAP and Pattern EAP *Chen Tian, University of Maryland, College Park*

Though scoring based on summed-score is more practical, the trade-off is decreased measurement precision. This study explores factors influencing the relative precision between summed-score EAP and pattern EAP, informing researchers the properties of EAP scores and how the relative precision covaries with test length, item heterogeneity, and models adopted.

The Cumulative Sum Procedures for Evaluating Differential Speededness *Suhwa Han, The University of Texas at Austin; Hyeon-Ah Kang, University of Texas at Austin*

The study presents cumulative sum (CUSUM) procedures for evaluating constancy of speed in response time models. The efficacy of the proposed procedures is assessed in simulations and real data analysis in comparison with existing person-fit methods. The implications of detecting differential speededness are explained under the framework of joint model of speed and ability.

The Effects of Rater Severity and Rating Design on Ability Estimation *XINYUE LI, Penn State University*

Given the problem of rater effect and incomplete design in research programs involved human rating, the purpose of this research is to address how application and extension of Many-Facet-Rasch-Model can inform researchers, teachers, psychologists in need of decision making.

The Nature of Differential Item Functioning in Items from Automatic Item Generator *Eunbee Kim, Georgia Institute of Technology; Susan Embretson, Professor*

This research tested whether items generated by an automatic item generator operate differently across genders and ethnicities. The effect sizes of DIF in generated items were negligible, comparable to that of classical operational items. For gender, the direction of item bias evened out, indicating no specific favor toward one gender.

The Psychometric Properties of the Chinese version of the Beck Depression Inventory-II with primary teachers *xiuna wang, Beijing Normal University*

We evaluated the psychometric properties of the Chinese version of the Beck Depression Inventory-II (BDI-II), which were responded by 1501 primary teachers in China, including the analysis of reliability, validity and measurement invariance with multiple-group confirmatory factor analysis. Results indicated that the Chinese version of the BDI-II has good reliability and validity, and can therefore be used to screen for depressive symptoms.

Using Odds Ratios to Detect Differential Item Functioning in a Computerized Adaptive Testing Environment *Siyu Wan, University of Massachusetts Amherst; Craig S. Wells, University of Massachusetts Amherst*

This study adopted a new proposed differential item functioning detection method by using Odds Ratio (OR) to the Computerized Adaptive Testing (CAT) Environment. The performance of OR method was compared to Logistic Regression (LR) and Mantel-Haenszel (MH) in CAT environment.

Utilizing Response Time in Omitted Response Treatment of Computer-Based Test *Haimiao Yuan, The University of Iowa*

A new non-response item scoring rule is proposed based on examinee's response time under the 4PLRT model. Several time thresholds were generated based on each examinee's least amount of time of being able to reach his/her maximum probability of answering an item correctly.

National Council on Measurement - ITEMS Module GSIC 2 *Andre Alexander Rupp, Educational Testing Service (ETS)*

Learn about your opportunity to publish in the ITEMS series. The ITEMS portal is your entry point into a world of learning in educational measurement and assessment. ITEMS modules are its centerpiece, which are short self-contained lessons with various supporting resources that facilitate self-guided learning, team-based discussions, as well as professional instruction.

099. NCME and AERA Division D Joint Reception

6:30 to 8:00 pm

Hilton San Francisco Union Square: Salon B

NCME and AERA Division D Joint Reception, all are welcome

100. NCME Breakfast

8:00 to 10:00 am

Hilton San Francisco Union Square: Salon B

NCME Breakfast

Awards Ceremony

Business Meeting

Presidential Address: *Yell Fire! Psychometricians in the Hands of an Angry Mob*

Stephen G. Sireci, University of Massachusetts Amherst

102. ** Invited Session ** Are you there NCME? It's me, Stakeholder.

Coordinated Session – Panel Discussion *12:25 to 1:55 pm*

Hilton San Francisco Union Square: Continental 4

A key aim of the National Council on Measurement in Education (NCME) is to serve as a valuable resource for diverse stakeholders (organizations, government agencies, individuals) in their use of educational assessments. Over the past year, the NCME Outreach Committee has been engaged in identifying ways in which NCME may reach its organizational mission through strategic outreach with local educational providers, agencies, and professional organizations. Towards this end, this session will offer invited panelists from across diverse roles in educational assessment to share their perspectives regarding the ways they feel the organization and its membership of measurement professionals can extend its outreach to support appropriate educational assessment use. The symposium will be moderated to promote honest dialogue among the panelists while providing feedback to NCME and measurement professionals, both as individuals and as a field about how we can improve practice to benefit key stakeholder groups.

Presenters:

Leigh Bennett, Loudoun County Public Schools
Shelley Loving-Ryder, Virginia Department of Education
Deborah Spitz, Office of Elementary and Secondary Education, U.S. Department of Education
Dorothea Anagnostopoulos, University of Connecticut
Thomas Philip, University of California, Berkeley

Session Organizer:

Tracey R Hembry, Alpine Testing Solutions

103. Advancing Multidimensional Science Assessment Design for Large-scale and Classroom Use

Coordinated Session *12:25 to 1:55 pm*

Hilton San Francisco Union Square: Imperial Ballroom A

Presenters will share the goals, progress, and national significance of the Strengthening Claims-based Interpretations and Uses of Large-scale Science Assessment Scores (SCILLSS) project funded through the US Department of Education's Enhanced Assessment Grants (EAG) program. SCILLSS brings together a consortium of three states, four organizations, and a panel of experts to strengthen the knowledge base among state and local educators for using principled-design approaches to design quality science assessments that generate meaningful and useful scores, and to establish a means for connecting statewide assessment results with classroom assessments and student work samples in a complementary system. Presenters will share how SCILLSS partners are applying current research, theory, and best practice to establish replicable and scalable principled-design tools that state and local educators can use to clarify and strengthen the connection between statewide assessments, local assessments, and classroom instruction, enabling all stakeholders to derive maximum meaning and utility from assessment scores.

Participants:

Ensuring Rigor and Strengthening Score Meaning in State and Local Assessment Systems *Ellen Forte, edCount, LLC*

A Principled-Design Approach for Creating Multi-Dimensional Large-Scale Science Assessments *Daisy Rutstein, SRI International*

A Principled-Design Approach for Creating Multi-Dimensional Classroom Science Assessments *Charlene Turner, edCount, LLC*

State Implementation of SCILLSS Resources: A User's Perspective *Rhonda True, Nebraska Department of Education*

Organizer:

Erin A Buchanan, edCount, LLC

Discussant:

Elizabeth Summers, edCount, LLC

104. Insights on Text-to-Speech as a Universal Design Feature: NAEP Mathematics Process Data

Coordinated Session *12:25 to 1:55 pm*

Hilton San Francisco Union Square: Imperial Ballroom B

The changing landscape of learning and assessment has led large scale assessments to transition to digitally-based assessments (DBAs). NAEP DBAs are designed to improve accessibility for all students via universal design (UD) principles, in addition to providing numerous tools in the digital assessment platform. One of NAEP's UD features is the text-to-speech (TTS) tool, which is a widely used accommodation in both large-scale and state-level assessments. TTS provides a delivery mechanism for audio presentation that removes the need to read materials. In DBAs, all students can select the TTS button and listen to written parts of any question in mathematics. Text-to-speech as an accommodation has a long history in the field of education; yet, there is no research on TTS as a UD feature conducted using process data to our knowledge. This symposium features three studies investigating the relationship between students' TTS tool use, characteristics of read-aloud sentences selected, and student performance. The first study examines use of TTS by student accommodation status. The second study examines the effect of TTS use by accommodation group. The third study explores features of selected read-aloud sentences. This symposium will contribute to the current literature on read-aloud functions, online testing, and item development.

Participants:

Evaluation of the Text-to-Speech (TTS) Tool Using NAEP Process Data *Soo Lee, American Institutes for Research*

Exploration of TTS Using the Differential Boost Framework *Juanita Hicks, American Institutes for Research*

Exploring Read-Aloud Sentences to Understand Learners' TTS Use Tendency *Ruhan Circi, AIR*

Session Organizer:

Juanita Hicks, American Institutes for Research

Discussant:

Heather Buzick, Educational Testing Service

105. Probabilistic Graphical Models for Writing Process Data

Coordinated Session *12:25 to 1:55 pm*

Hilton San Francisco Union Square: Yosemite A

Most of the research literature around writing process, in the field of educational measurement, has primarily focused on feature development and exploring the applications of the writing process features for the purpose of essay scoring, gaining new knowledge about writing and different writing tasks, understanding individual or subgroup differences, improving test validity, as well as profiling of students' writing patterns or writing styles. Less work has been targeted at statistical modeling of the writing features or the overall writing process. In this symposium, we present four papers around using probabilistic graphical models to understand a writer's or groups of writers' composition process. Despite of its increasing popularity in the psychometrics literature, most development in graphical modeling has been centered on the modeling item response; its application in treating timing and process data is rarely discussed. We hope this symposium will not only demonstrate the utility of this type of models in treating writing process data, but also inspire greater interests in this line of research among the measurement community when dealing with timing and process data.

Participants:

Effects of Scenario-Based Assessment on Students' Writing Processes *Hongwen Guo, Educational Testing Service; Mo Zhang, ETS; Paul Deane, ETS; Randy E Bennett, ETS*

Examine the Prompt Effects on Composition Process Using Mixed Markov Process Model *Mo Zhang, ETS; Xiang Liu, Educational Testing Service; Hongwen Guo, Educational Testing Service*

Consistency and Predictive Power of Keystroke Feature Variables over Time *Mengxiao Zhu, Educational Testing Service; Xiang Liu, Educational Testing Service*

Bayesian Nonparametric Graphical Model for Writing Features *Xiang Liu, Educational Testing Service*

Session Organizer:

Xiang Liu, Educational Testing Service

Chair:

Hongwen Guo, Educational Testing Service

Discussant:

Matthew Johnson, Educational Testing Service

106. Demonstrating the Utility of Generalizability Theory across Heterogeneous Assessments

Coordinated Session *12:25 to 1:55 pm*

Hilton San Francisco Union Square: Yosemite B

Validity is the paramount consideration for psychometricians, but valid score interpretations are impossible without reliable scores. Common estimates of reliability like coefficient alpha provide little indication of the sources of measurement error. Assessment developers will be better informed using generalizability (G) theory, a comprehensive approach to reliability estimation. Moreover, G Theory can be used to predict the impact on reliability for different number of measurement replications (e.g., more items). This collection of studies provides an illustration of the usefulness of G Theory, especially during the assessment development phase when changes can be made. G Theory is used to evaluate the reliability of a teacher observation tool, an assessment providing formative feedback to educational leaders, and a game-based assessment of conscientiousness. Each study leverages different study designs to evaluate the reliability of the current assessments and estimate the change in reliability for different numbers of items, raters, etc. These three examples demonstrate the versatility of G Theory across different domains of educational assessment. In keeping with the theme of the conference, educational experts outside of psychometrics will provide important contextual information about the purpose and development of the assessments.

Presenter:

Katherine McKnight, RTI International

Participants:

Evaluating Reliability of the Lead Learner Teacher Observation Tool *Katherine McKnight, RTI International; Sonya Powers, RTI International*

Reliability of a 360 Assessment for the Professional Development of Educational Leaders *Sonya Powers, RTI International; Rakhee Patel, The Broad Center; Nitya Venkateswaran, RTI International; David Silver, IMPAQ International; Andrea Foggy-Paxton, The Broad Center*

Reliability of a Game-Based Assessment of Conscientiousness for Middle School Students *Sonya Powers, RTI International; Carmen Strigel, RTI International; Sarah Pouezevara, RTI International; Katherine McKnight, RTI International; Greg Moore, Independent Consultant*

Session Organizer:

Sonya Powers, RTI International

Discussant:

Robert Brennan, Professor Emeritus, University of Iowa, Center for Advanced Studies in Measurement and Assessment

107. Fairness in Educational Measurement: The Issue of Measurement Invariance

Coordinated Session *12:25 to 1:55 pm*

Hilton San Francisco Union Square: Yosemite C

In education measurement researchers are often interested in regressing the ability underlying educational tests on predictor variables like gender and socioeconomic status. By doing so, one needs to assume measurement invariance. This assumption holds that there is no differential item functioning such that the measurement model parameters are the same across the levels of the predictor variables. In the case of categorical predictor variables with a limited number of levels (e.g., gender) and dichotomous item scores, well-established statistical tools are available to test for measurement invariance or differential item functioning. However, there are many practical situations in educational measurement where these well-established statistical tools do not suffice. In the present symposium, we focus on situations that are often encountered in educational measurement practice, but in which the current measurement invariance methodology is suboptimal or not well-established yet. We will outline the challenges associated with these practical situations and we will propose new methods to test for measurement invariance. All talks are intended to give concrete advice on how to test for measurement invariance in practical educational measurement settings.

Participants:

Measurement Invariance with ordinal data: a comparison between CFA and IRT *Damiano D'Urso, University of Tilburg; Jesper Tijmstra, Tilburg University; Kim De Roover, Tilburg University; Jeroen Vermunt, Tilburg University*

Mixture multigroup factor analysis for unraveling measurement non-invariance across many groups *Kim De Roover, Tilburg University; Eva Ceulemans, KU Leuven; Jeroen Vermunt, Tilburg University*

Multi-group scalar invariance inferences from multilevel factor models *Jingdan Zhu, Ohio State University; Paul De Boeck, Ohio State University*

Latent Markov Factor Analysis for Exploring Measurement Model Changes over Time *Leonie Vogelsmeier, Tilburg University; Jeroen Vermunt, Tilburg University; Kim De Roover, Tilburg University*

A Semi-Parametric Approach to Test for Measurement Invariance across a Continuous Variable *Dylan Molenaar, University of Amsterdam*

Session Organizer:

Dylan Molenaar, University of Amsterdam

108. New Research Findings on Understanding and Managing Test-Taking Disengagement

Coordinated Session *12:25 to 1:55 pm*

Hilton San Francisco Union Square: Continental 1

Over the past two decades, there has been a growing interest in the impact of disengaged test taking on test performance and score validity. This problem is most often observed in low-stakes testing, in which test takers perceive few, if any, consequences associated with their performance. The research on disengagement has followed several general themes: measuring test-taking engagement and detecting disengagement, estimating the amount of score distortion due to disengagement and adjusting scores for its impact, uncovering the psychological dynamics of disengagement, and understanding the types of items and testing conditions that elicit disengaged test taking. The papers in this session, provided by leading researchers in the field, address several of these themes. The first two papers investigate the utility of a method for statistically adjusting scores for the effects of rapid-guessing behavior. The third paper looks at how specific design features of technology-enhanced items influence test-taking engagement. The final two papers investigate the nature of disengagement and its association with several psychological and situational variables.

Participants:

Aggregate-Level Ability Estimation Accuracy Under Varying Non-effortful Responding Types and Rates *Joseph A. Rios, University of Minnesota; Chelsey Legacy, University of Minnesota*

Using Retest Data to Evaluate and Improve Effort-Moderated Scoring *Steven L. Wise, NWEA; Megan R. Kuhfeld, NWEA*

How to Design a Drag-and-Drop Item for Motivated Responding *Blair Lehman, ETS; Burcu Arslan, ETS*

Test Value and Emotions: Predicting Examinee Effort and Performance on Low-Stakes Tests *Paulius Satkus, James Madison University; Sara J. Finney, James Madison University*

Do Students Rapidly Guess Repeatedly Over Time? A Longitudinal Analysis *James Soland, NWEA; Megan R. Kuhfeld, NWEA*

Session Organizer:

Steven L. Wise, NWEA

Chair:

Dena Pastor, James Madison University

109. Automation in Test Development and Scoring

Paper Session *12:25 to 1:55 pm*

Hilton San Francisco Union Square: Continental 2

Participants:

An Optimal Composite Estimator Using a Human Score and a Model-yielded Score *Jiahe Qian, Educational Testing Service*

Automated Evaluating for Main Idea in Chinese Composition Based on Concept Maps *Tao Xin, Beijing Normal University, PRC.; liping Yang, Beijing Normal University*

This study proposes a method to extract text features using concept maps to assess the main idea in Chinese composition, and provides feedback for students. The results report 0.82 quadratic weighted Kappa and 91% exact plus adjacent agreement between the predicted scores and the human scores.

Automated Reading Item Generation Using Topic Modelling Apporach *Jinnie Shin, University of Alberta; Mark J. Gierl, University of Alberta; Hollis Lai, University of Alberta; Donna Matovinovic, ACT Inc.*

A three-stage topic modelling approach is developed to understand the latent topic structure of a fictional text. We demonstrated the current system's benefits by generating multiple-choice reading inference items automatically. The results indicated that our system could investigate the topic structure of the Harry Potter novel while creating items systematically.

Automated short answer scoring using an ensemble of neural networks *Christopher M Ormerod, American Institutes for Research; Susan Lottridge, AIR; Balaji For Kodeswaran, American Institutes for Research; Paul For Vanwamelen, American Institutes for Research; Milan For Patel, American Institutes for Research; Amy For Harris, American Institutes for Research*

We apply a neural network-based engine to grade short constructed responses to a large suite of questions from a national assessment program. We evaluate the performance of the engine, discuss the issues in large-scale implementation and explore the effects of using different semantic word vectors.

Evaluating Subpopulation Invariance in Automated Essay Scoring *Wei Wang, Educational Testing Service; Jing Miao, Educational Testing Service; Neil Dorans, Educational Testing Service; Chi-wen Liao, Educational Testing Service; Chen Li, Educational Testing Service*

Population invariance is an important measure in evaluating test fairness. The proposed study examines subpopulation invariance in automated essay scoring using data from a testing program. The subgroups of interest are genders, ethnicities, and their combinations.

Chair:

Tyler Holmes Matta, Pearson

Discussant:

James Burns Olsen, Renaissance

110. Cognitive Diagnostic Models #2

Paper Session *12:25 to 1:55 pm*

Hilton San Francisco Union Square: Continental 3

Participants:

A Bayesian Limited Information Approach to Diagnostic Classification Model-Data Fit *Catherine Elizabeth Mintz, University of Iowa; Jonathan Templin, University of Iowa; Jihong Zhang, University of Iowa*

This study investigates an explicitly Bayesian PPMC method using a diagnostic classification model (DCM) as example. A Bayesian limited-information saturated model was fit and used as a comparison to a hypothesized DCM. Results suggest the Bayesian limited information saturated model approach is accurate and superior to traditional PPMC methods.

A Note on Model Fit Evaluation of Cognitive Diagnosis Models *Sean Oliver Mojica Escalante, University of the Philippines Diliman; Kevin Carl Santos, University of the Philippines; Iris Ivy M. Gauran, University of the Philippines*

This study evaluates the sensitivity of existing absolute fit statistics in cognitive diagnosis models on item quality, and adapts a new fit statistic from item response theory. Findings reveal that the proposed statistic provides superior power and type I error rates when items are of at least medium quality.

Context Matters: A Comparison of Empirical CDM Analyses Involving Two Different Q-Matrices *Hartono Tjoe, The Pennsylvania State University; Qianru Liang, The University of Hong Kong; Jimmy de la Torre, Hong Kong University*

This study investigates how Q-matrices developed using two different curricula (i.e., US and Hong Kong) affect model-data fit and inferences about examinees' mastery profiles. Analyses of proportional reasoning test data collected in Hong Kong show that, although the two Q-matrices provide similar model-data fits, their examinee classifications are dramatically different.

Estimating Cognitive Diagnosis Models in Small Samples: Bayesian Modal Estimation and Monotonic Constraints *ZHEHAN JIANG, Baylor College of Medicine; Wenchao Ma, The University of Alabama*

Cognitive diagnosis models have been criticized for limited utility for small samples. We proposed to use Bayes Modal estimation with monotonic constraints to improve person classification accuracy. Simulation study was conducted to compare GDINA parameter recovery using traditional and the proposed algorithms under varied conditions, the results favored latter approach.

Multilevel Mixture Independent and Higher-Order Cognitive Diagnosis Model *Kuan Xing, University of Illinois at Chicago; QIAO LIN, University of Illinois at Chicago; Yoon Soo Park, University of Illinois at Chicago*

Prior research on cognitive diagnosis model (CDM) has focused on explanatory factors and multilevel analysis to extend CDM on student achievement data. This study proposes a family of multilevel independent and higher-order CDM, allowing mixture extensions. Real-world data analysis and simulation study are conducted. Applications of the model are discussed.

Chair:

Masha Bertling, Harvard University

Discussant:

Andre Alexander Rupp, Educational Testing Service (ETS)

111. Equity and Fairness in Testing

Paper Session *12:25 to 1:55 pm*

Hilton San Francisco Union Square: Franciscan Ballroom A

Participants:

Countering the Deficit Narrative in Quantitative Educational Research *Michael Russell, Boston College; Carly Oddleifson, Boston College; Lawrence Kaplan, Boston College*

This paper presents findings from an analysis of 200 peer-reviewed articles that present findings from quantitative research in which race is employed as a variable. The paper focuses specifically on the extent to which the presentation of findings in these articles contributes to deficit narratives and offers approaches to presenting such findings in a manner that supports anti-racist narratives.

How does digital assessment impact item performance among disadvantaged test-takers? *Elena Mariani, Pearson; Kevin Mason, Pearson*

We combine administrative data with assessment information to examine the relative performance of similar items in paper-based and computer-based tests among historically disadvantaged test-takers, defined in terms of ethnicity, economic background and learning difficulty. The aim is to establish how a transition to computer-based assessment would impact groups of learners.

Should Psychometricians Make Claims About Test Fairness? *Daniel Katz, University of California, Santa Barbara; Anthony Clairmont, University of California, Santa Barbara*

The Standards for Educational and Psychological Testing (AERA, NCME, APA, 2014) have a chapter on test fairness that neglects to define fairness. We present common definitions of fairness and question whether these definitions of fairness are probed by psychometrics. We map test results across New York City as an example.

The Interpretability and Use of Accountability Equity Scores for Improvement *Nikole Gregg, James Madison University; Brian Gong, Center for Assessment*

One of the intentions of ESSA is to improve equity. We consider how an accountability indicator can be used and interpreted to inform equity interventions. We investigate a state's equity measure and propose a new form of the equity measure that is sensitive, stable, and aligns with an intended purpose.

Using Multiple Measures to Mitigate the Impact of Selection Bias *Andrew Jones, American Board of Surgery; Jason P Kopp, American Board of Surgery; Beatriz Ibanez, American Board of Surgery; Derek Sauder, James Madison University*

Selection bias, wherein the use of test scores for classification results in non-equivalent classification errors between subgroups, warrants more attention in the psychometric literature. This study examines how the use of multiple measures may mitigate the effects of selection bias under certain conditions.

Chair:

Rudolf Debelak, University of Zurich

Discussant:

Michael Clifford Rodriguez, University of Minnesota

112. CAT Design and Application

Paper Session *12:25 to 1:55 pm*

Hilton San Francisco Union Square: Franciscan Ballroom C

Participants:

Comparability of Scoring Incomplete Linear and Computerized Adaptive Tests *Yu Fang, Law School Admission Council; Deborah Harris, University of Iowa; Troy Chen, ACT, Inc.; Yang Lu, ACT, Inc.*

In CAT settings, it is challenging to provide fair scores to examinees with incomplete tests. To address this issue, scoring approaches using a penalty function, prediction or item simulation have been proposed. This study investigates score estimates generated from these methods for incomplete tests in linear and CAT test settings.

Drift Analysis for Polytomous Items in a Computer Adaptive Test *CHANGJIANG WANG, Pearson; David Shin, Pearson*

We propose two pseudo-count-based methods for examining drift in polytomous items in CAT, an extension of similar methods for dichotomous items. The detection power and the misclassification of each method will be documented. The results can be used as guidance for assessment practitioners to safeguard the healthiness of CAT banks.

Evaluating Grade-Level Item Pools for Large-Scale Computerized Adaptive Tests *Xueming Li, NWEA; Patrick Meyer, NWEA*

This study examines measurement precision, item exposure rates, and the depth of item pools under various grade-level restrictions. The type of item pool and type of simulee were manipulated in the simulations. Results will support test development with these item pools and provide recommendations for developing grade-level items.

Grid Multiclassification Adaptive Classification Testing with Multidimensional Polytomous Items *Zhuoran Wang, Prometric; Chun Wang, University of Washington; David J Weiss, University of Minnesota Twin Cities*

Termination criteria were proposed for grid multiclassification adaptive classification testing (ACT) which aims at allocating each examinee into one of the grids encircled by cutoff scores(lines/surfaced) along different dimensions. A simulation study shows grid multiclassification ACT is more efficient than measurement CAT-based classification, especially when considerable dimension correlation exits.

Improving Ability Estimation with Collateral Information in an Adaptive Battery *Qing Xie, ETS; Deborah Harris, University of Iowa*

This study compares different collateral information methods in ability estimation under the unidimensional and multidimensional CAT frameworks and examines the impact of item pool characteristics and stopping rules on performance of the CI in a variable-length adaptive battery with content constraints and item exposure control.

Proficiency Estimation in Computerized Adaptive Testing Using a Locally Objective Prior *Can Shao, Curriculum Associates; David Thissen, University of North Carolina at Chapel Hill; Li Cai, University of California—Los Angeles; Kevin Cappaert, Curriculum Associates; Michael Edwards, Arizona State University; Yawei Shen, The University of Georgia*

This paper proposes a new proficiency estimation method for computerized adaptive testing (CAT): the expected a posteriori (EAP) estimator with a locally objective prior. The method draws from Jeffreys' (1946) method and may produce less shrinkage for high and low performing test takers than EAP with normal population-based priors.

Chair:

Janet Mee, National Board of Medical Examiners

Discussant:

Laurie Laughlin Davis, Curriculum Associates

101. Electronic Board. Sunday 12:25

Electronic Board Session
12:25 to 1:55 pm
Hilton San Francisco Union Square: Salon A
Participants:

Understanding and Illustrating Use of Test Accessibility Supports *Dukjae Lee, University of Massachusetts Amherst; Stephen Sireci, University of Massachusetts Amherst; Cara Laitusis, Educational Testing Service; Heather Buzick, Educational Testing Service; Mina Lee, University of Massachusetts Amherst; Michelle Center, California Department of Education*

To promote fairness and access in assessment, accessibility supports are offered to students with disabilities and English learners. We analyzed data from a large state and found virtually no unapproved use, but there were many approved students who did not use supports. Using data visualization procedures, aberrant districts were identified.

Accuracy of Scale Linking Under Two Conflicting IRT Paradigms *Jaime Malatesta, Center for Advanced Studies in Measurement and Assessment; Kuo-Feng Chang, University of Iowa; Won-Chan Lee, University of Iowa*

This paper has three main objectives: 1) discuss the differences between the 1PL model and Rasch model with respect to scaling, 2) evaluate the accuracy of several IRT scale linking methods for each model under various conditions, and 3) compare results from this study with those of previous studies.

A Monte Carlo Simulation on Group-Lasso, -Enet, and -Mnet *Jin Eun Yoo, Korea National University of Education; Minjeong Rho, Korea National University of Education*

The performance of group Lasso, group Enet, and group Mnet was evaluated with missing data techniques and variable selection criteria in a Monte Carlo simulation. Each condition had 100 replications of 340 variables and 2,000 observations. The evaluation criteria included accuracy, kappa, and the number of true variables unselected.

Bias in Estimation of the C Parameter *Briana Hennessy, University of Connecticut; Eric Loken, University of Connecticut*

Estimation of the c parameter for tests modeled with the 3PL model can be challenging. This study investigates c parameter estimation and finds downward bias. We also report 95% coverage as a function of the a and b parameters.

Comparison of Fixed Items and Fixed Persons Concurrent Calibrations for Small Samples *M. Kuzey Bilir, Pearson; Steven Fitzpatrick, Pearson; Ou Zhang, Pearson*

This study evaluates accuracy and stability of item parameters for Rasch/PCM models, and investigates RMSE for two equating designs; concurrent calibration by fixing a) common item parameters, b) examinee ability estimates which enables to double sample size for common items, and therefore likely reduces RMSE of equating when sample size is very low.

Impact of Guessing on Item Response Theory Vertical Scaling in Small-Scale Assessment *Lanrong Li, Florida State University; Insu Paek, Florida State University*

We investigated the impact of fitting different item response theory models to item responses affected by guessing on the results of vertical scaling in small-scale assessment. Our results showed that fitting the 1-PL model generally led to smaller bias in person-related parameter estimates than fitting the 2-PL model.

Inequalities in Access to Educational Opportunities: Investigation of PISA Dataset Using Multilevel-IRT *Jayashri Srinivasan, UCLA*

Increasingly, PISA tests are influencing education policies across multiple countries. This study makes use of a multilevel IRT framework to examine the issue of inequalities in access to various instructional practices, and assess measurement invariance across rural and urban regions using the PISA 2009 dataset.

Investigating Multidimensionality in the Comprehensive Assessment of Outcomes in Statistics *Vimal V Rao, University of Minnesota*

The Comprehensive Assessment of Outcomes in Statistics measures students' conceptual understanding of learning outcomes in statistics. I apply multidimensional item response theory methods to garner validity evidence supporting subscore utilization. Results show statistical significance for multidimensionality yet fail to satisfy practically significant thresholds. Subscores use is not recommended.

Investigating Repeater Subgroup Effects on Score Equating *Jiawen Zhou, Educational Testing Service; Yi Cao, Educational Testing Service*

Retesting policy is common for most testing programs. The purpose of this study is to explore the impact of including or excluding certain repeater subgroup(s) on score equating using empirical data of an authentic licensure testing.

Investigating the Impact of Treatments for Repeating Response upon Item Parameter Calibration *Mingjia Ma, University of Iowa; Zhongmin Cui, ACT, Inc.*

Much research has been focusing on the treatment of missing data with respect of item parameter estimates; this research tries to investigate the impact of extreme long repeating responses identified by Cz index upon standard errors of item parameter estimates. It's expected to obtain more stable parameter estimates.

IRT Modeling with the Complementary Log-Log Link Function *Hyejin Shim, University of Missouri-Columbia; Wes Bonifay, University of Missouri*

This study demonstrates how application of the Complementary Log-Log (CLL) link function in IRT modeling yields reliable and interpretable parameter estimates in the presence of zero-inflated response data. Further, this work shows that in certain cases, the simple 1-parameter CLL model is comparable to a more complex 2-parameter model.

Item Drift Evaluation Approaches in Large-scale Assessment *Xiaoxin Wei, AIR; Tao Jiang, American Institutes for Research; Yuan Hong, Americian Institutes for Research*

In large-scale assessments, item parameter drift jeopardizes test validity and compromises ability estimation. This study compared a fit statistic approach and a linking approach to evaluate drift using simulated and empirical data. The result suggested each approach has its advantages and disadvantages that need to be considered in practice.

Item Response Theory Parameter Recovery in Mixed Format Test Using FlexMIRT *huan liu, The University of Iowa; Won-Chan Lee,*

University of Iowa

The main purpose of this study was to investigate the performance of commercial software program flexMIRT on item parameter recovery under mixed format test conditions. Three factors of investigation were considered in this research: model combination, sample size and item number. The preliminary results were presented.

Model Selection with Bayesian Confirmatory Multidimensional Item Response Theory Models *Ken Fujimoto, Loyola University Chicago; Carl F Falk, McGill University*

Through simulations, we examined how well five Bayesian model fit indices could identify the correct multidimensional item response theory (MIRT) model. The results revealed that the Pareto-smoothed-importance-sampling-based leave-one-out cross-validation consistently identified the correct MIRT model in all conditions, whereas the deviance information criterion did so in only some conditions.

Modeling Latent Ability Change in a Longitudinal Assessment: a MIRT Approach *Matthew John Davidson, University of Washington; Fen Fan, NCCPA; Drew Dallas, NCCPA; Joshua Goodman, NCCPA; John Weir, NCCPA*

Response data from a longitudinal re-certification assessment is analyzed using a MIRT model adapted for estimating growth parameters, to investigate whether theta changes between testing occasions. Results suggest that theta decreases over time, with important implications for modeling longitudinal response data.

Optimal Design for Online Calibration with Polytomous Items *Hao Ren, Pearson; Seung Choi, The University of Texas at Austin; Wim J van der Linden, University of Twente*

Interest in adaptive testing with polytomous items has grown considerably, and hence the need of item field testing. Efficient optimal design for online calibration seems to be an attractive approach to pursue. This study extended the optimal design to handle polytomous items and extensive simulation results will be presented.

Penalized Estimation of Polytomous IRT models *Alex Brodersen, University of Notre Dame*

Many IRT models are special cases of the divide-by-total version of the nominal response model. If sample size is not sufficient or there is model selection uncertainty, these models may provide unstable estimation. This talk provides alternative penalization schemes that provide additional stability in item parameter estimation.

Recovering from Suboptimal Precalibration in a Low-Stakes, Computerized-Adaptive, Diagnostic Assessment System *Anthony Albano, University of California, Davis; Josine Verhagen, Kidaptive*

When developing online learning products, practical constraints often limit the feasibility of standard procedures for IRT parameter estimation. This study examines how assessment results are impacted by the use of 1) expert ratings of item difficulty in place of IRT precalibrated values and 2) recalibration methods based on multi-unidimensional models.

The Relationship between IRT Difficulty and Discrimination *Sandra M Botha, University of Massachusetts - Amherst; Sandip Sinharay, ETS; Matthew Johnson, Educational Testing Service*

This study focused on the investigation of the empirical relationship between item difficulty and discrimination in an IRT framework. Results show negative correlations for the 2PL model and positive correlations for the 3PL model, though the magnitude varied for different tests. Implications to practice and future directions are discussed.

Two Parameter Multilevel IRT Model with Heterogeneous Within-Group Variances *Yusuf Kara, Southern Methodist University; Akihito Kamata, Southern Methodist University*

This study proposes a two-parameter multilevel IRT model with heterogeneous within-group variances. We introduce the model equations along with parameter estimation procedure adopting the Bayesian approach.

Using NAEP Science Achievement Test Data to Explore the Dimensional Structure of a Complex Large-Scale Assessment Using Multidimensional Item Response Theory *ARIFE KART ARSLAN, Duquesne University; Cindy Walker, Duquesne University; Ömer Kutlu, Ankara University, Turkey*

The model-data fit, item-model fit, and the dimensionality of the NAEP Assessment are determined by examining the parameters in item level with overall goodness of fit indices both separate and simultaneously calibration. The combination of the item type, the science content areas and the cognitive skills has been taken into consideration within the scope of the UIRT, MIRT and BF models.

114. Psychometrics is Dead - Long Live Psychometrics: Measurement Still Matters.

Coordinated Session *2:15 to 4:15 pm*

Hilton San Francisco Union Square: Continental 4

Measurement science, like many other disciplines, has been, and continues to be, severely impacted by the ushering of the information age. Traditional stalwarts, such as episodically administered fixed form MCQ exams, are being challenged in an era when crystalized intelligence-based skills may no longer completely mirror the breadth of abilities expected in order meet the challenges and opportunities that define current times. Skills such as critical thinking, problem solving, creativity, adaptability and information literacy are being heralded as critical areas for development, and by extension, assessment. The aim of this coordinated session is to outline four areas of research that highlight how measurement science is evolving to continue to be relevant in this fast-changing landscape. The papers in this session will summarize work in the following four areas as a means of highlighting how traditional psychometrics is, and has been morphing, both in terms of its growing multidisciplinary nature and shifts in philosophical underpinnings: (1) computational psychometrics, (2) the development of task models, (3) automated item generation and (4) adaptive practice systems. It is our hope that this session will generate rich discussion on how our field can continue to be responsive and critical to a myriad of stakeholders.

Participants:

Computational Psychometrics for Learning & Assessment Systems *Alina von Davier, ACTNext*

No More Items: On the Principled Design of Task-Model Families *Richard Luecht, University of North Carolina*

The Evolving Needs of Item Generation *Hollis Lai, University of Alberta*

Making Measurement Matter: Computerized Adaptive Practicing *Han van der Maas, University of Amsterdam*

Session Organizer:

Andre De Champlain, Medical Council of Canada

Chair:

Andre De Champlain, Medical Council of Canada

Discussant:

Andrew Maul, University of California - Santa Barbara

115. **Invited Session ** Assessing Indigenous Students: Co-Creating a Culturally Relevant & Sustaining Assessment System

Coordinated Session – Panel Discussion *2:15 to 4:15 pm*

Hilton San Francisco Union Square: Imperial Ballroom A

The NCME Diversity Issues in Testing Committee is pleased to offer an invited panel session at the NCME 2020 conference in San Francisco focused on assessment issues affecting indigenous students. In the invited panel session that was held at the 2019 NCME conference in Toronto, discussion focused on how to make assessments more equitable for students of color. This session extends that discussion with a focus on the equitable assessment of indigenous students, specifically, by naming and addressing the unique challenges these students face within the context of traditional systems of assessment. How can we help indigenous students succeed in an educational system that has failed them?

Presenters:

Madhabi Chatterji, Columbia University, Teachers College
Kerry Englert, Seneca Consulting
Kalehua Krug, Hawaii Department of Education
Pohai Kukea Shultz, University of Hawaii
Sharon Nelson-Barber, WestEd

Session Organizers:

Leanne R. Ketterlin-Geller, Southern Methodist University
Katherine Schlatter, Columbia University

116. Analyzing Students' Process Data in a Science Game-Based Assessment

Coordinated Session *2:15 to 4:15 pm*

Hilton San Francisco Union Square: Imperial Ballroom B

Digital game-based assessments (DGBA) are tools that may measure process-based competencies because they mimic real-world situations and capture evidence of knowledge and skills in an engaging environment designed to reduce anxieties related to testing. The DGBA program creates evidence-trace files, list of students' actions throughout the task, which may be analyzed for evidence of students' proficiency on process-based competencies related to student problem solving. The four papers included in this symposium present different techniques, which complemented one another, for analyzing the evidence-trace files of a DGBA called Raging Skies. The use of multiple techniques was to ensure that valid claims about achievement proficiency of competencies were made.

Participants:

Analyzing Student Process Data with Bayesian Knowledge Tracing and Dynamic Bayesian Network *Ying Cui, University of Alberta; Man-Wai Chu, University of Calgary; Fu Chen, University of Alberta; Qi Guo, University of Alberta*

Utilizing Game Analytics to Inform Digital Game-Based Assessment Design *Fu Chen, University of Alberta; Qi Guo, University of Alberta; Ying Cui, University of Alberta; Man-Wai Chu, University of Calgary*

Validity of Process-Based Competency Outcome Claims Using Think-a-Loud Data and Evidence-Trace Files *Man-Wai Chu, University of Calgary; Ying Cui, University of Alberta; Mahnaz Shojaee, University of Alberta; Maryam Hachem, University of Calgary; Qi Guo, University of Alberta; Fu Chen, University of Alberta*

Detection of Aberrant Response Patterns Using Discrete Variational Autoencoder *Qi Guo, University of Alberta; Fu Chen, University of Alberta; Man-Wai Chu, University of Calgary; Ying Cui, University of Alberta*

Session Organizer:

Man-Wai Chu, University of Calgary

Discussant:

Jimmy de la Torre, Hong Kong University

117. Graduate Training in Educational Measurement and Psychometrics: A Curricula Review of Graduate Programs in the U.S.

Coordinated Session – Panel Discussion *2:15 to 4:15 pm*

Hilton San Francisco Union Square: Yosemite A

A curriculum review of 135 graduate (masters & doctoral) programs in educational measurement, assessment, evaluation, psychometrics, and/or quantitative psychology in the United States was conducted to examine both the content and skills prioritized in graduate training in the field. The review consisted of a content analysis of the explicit program requirements/electives as well as interviews with a cross-section of current educational measurement professionals. In addition to required content, programs/program curricula were coded with respect to intellectual home (psychology v. education departments), region, level of program (M.A. v. EdD or PhD), total credits, number and rank of faculty. Patterns with respect to content variation will be presented followed by a facilitated discussion around measurement curriculum (what must we keep, what is missing, etc.). In sum this session will consider the question: What does it mean to be an educational measurement specialist?

Presenters:

Chad Buckendahl, ACS Ventures, LLC
Sara J. Finney, James Madison University
Matthew Madison, Clemson University
Kristen Huff, Curriculum Associates
Ye Tong, Pearson

Session Organizer:

Jennifer Randall, University of Massachusetts Amherst

Discussant:

Joseph A. Rios, University of Minnesota

118. Beyond Accommodations: Intentional Design Methods for Improving Validity When Assessing Diverse Populations

Coordinated Session *2:15 to 3:45 pm*

Hilton San Francisco Union Square: Yosemite B

Development of fair, accurate, and valid assessment solutions requires a grounded understanding of the intersection of students' abilities and challenges both aligned with and orthogonal to the learning objectives intended for measurement. Attempting to solve accessibility problems through retrofitted accommodations general fails for two reasons: (1) the tools and interfaces given to students are often limited in utility and don't provide the supports needed for students to fully demonstrate their construct-relevant knowledge, skills, and understandings, and (2) post-hoc accommodations can compromise intended item constructs. Instead, validity must be addressed throughout the test development process, using intentional processes to minimize the impact of students' construct-irrelevant abilities. Three approaches—use of explicit theories of action during test development, item twinning of science assessment items, and use of principled accessibility guidelines during test development—are described as means for providing these intentional processes to create large-scale assessment solutions that are fair, accurate, and valid for a wide range of learners, including those with disabilities.

Participants:

Improving Fairness: Theories of Action and Intentional Design Methods *Melissa L. Gholson, Educational Testing Service*

Equitable Assessment through Twinning *Danielle Guzman-Orth, Educational Testing Service; Cary Supalo, Educational Testing Service; Cinda Parton, WestEd*

Applying Universal Design Principles to the Accessible Test Development Process *Robert P Dolan, Diverse Learners Consulting; Cara Wojcik, CAST; Jenna Gravel, CAST; Allison Posey, CAST; Elizabeth Hartmann, Lasell College; Kimberly Ducharme, CAST; Jose Blackorby, CAST*

Session Organizer:

Robert P Dolan, Diverse Learners Consulting

Discussant:

Sheryl Lazarus, National Center on Educational Outcomes

119. Procedures for Establishing and Evaluating Linkages between Scores Collected in Different Modes

Coordinated Session *2:15 to 4:15 pm*

Hilton San Francisco Union Square: Yosemite C

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

Participants:

Design considerations for linking across modes *Yue Jia, Educational Testing Service; Nuo Xi, Educational Testing Service*

Common population linking method used in NAEP digital transitions *Nuo Xi, Educational Testing Service; Paul A Jewsbury, ETS*

IRT model extensions for modeling mode effects in PISA 2015 *Matthias von Davier, National Board of Medical Examiners; Lale Khorramdel, National Board of Medical Examiners*

Mode comparability studies in K-12 state testing *Katherine Castellano, Educational Testing Service*

Session Organizer:

Nuo Xi, Educational Testing Service

Discussants:

Neil Dorans, Educational Testing Service

Mary Pommerich, Defense Manpower Data Center

120. The Struggle is Real: Tackling Alignment Challenges in a Changing Assessment World

Coordinated Session – Panel Discussion *2:15 to 3:45 pm*

Hilton San Francisco Union Square: Continental 1

Alignment has served as a principal criterion in the validity evaluation of standards-based assessments. However, changes in the educational landscape including innovative approaches to the design of an educational system, more complex systems of standards, different types of measurement strategies, and alternative approaches to assessment design have strained the credibility of traditional alignment approaches. Unlike other areas of assessment design and evaluation, alignment methodology has had relatively less attention and scrutiny in professional settings and scholarly publications. In this session, panelists who work with educational assessment systems designing, conducting, and evaluating alignment studies will present and discuss such challenges, including: 1. Defining the purpose of the alignment evaluation, 2. Identifying the components of the system to be aligned (e.g., claims, standards, PLDs/ALDs, test form(s), score scales), 3.Determining the scope of the alignment study, 4.Determining the classes of evidence used for evaluating alignment (e.g., content, knowledge/skills, cognitive complexity, judgmental consistency), and 5.Interpreting the alignment results (e.g., weight attributed to alignment vs. other types of evidence, and whether the evaluation/interpretation should vary based on the claims). Session attendees will have the opportunity to ask questions of the panel and even share their own ideas.

Presenters:

Wayne J. Camara, ACT
Ellen Forte, edCount, LLC
Scott Marion, National Center for the Improvement of Educational Assessment

Session Organizer:

Susan Davis-Becker, ACS Ventures, LLC

Chair:

Thanos Patelis, Fordham University

121. State Testing

Paper Session *2:15 to 4:15 pm*

Hilton San Francisco Union Square: Continental 2

Participants:

A Systematic Review of Nationally Available Testing Accommodations for English Learners *Carlos Chavez, University of Minnesota; Joseph A. Rios, University of Minnesota; Samuel D Ihlenfeldt, University of Minnesota*

Despite a national priority to fairly assess the academic success of English learners (ELs), EL test accommodation practices remain understudied since the passage of ESSA. To address this limitation, this study conducted a systematic review of state accountability assessment program EL accommodation practices across all 50 states in the U.S.

Are They Trying? Motivation on a State-Mandated Assessment of Career Readiness *Jeffrey T. Steedle, ACT*

High school students can benefit from demonstrating career readiness on workplace skills assessments, but motivation is a concern in low-stakes test administrations. This study examined five indicators of motivation based on test data. Compared to adults testing under high-stakes conditions, more high school students were flagged for possible low motivation.

Open or Closed Directions for Multiple Response Items in K-12 Testing *Yufeng Chang Berry, Minnesota Department of Education; Angela Hochstetter, Minnesota Department of Eduation; Tony D Thompson, Pearson*

The effect of changing open directions (e.g., "Select all that apply") on multiple response items to closed directions (e.g. "Select n") on measurement properties was investigated. This study compares 66 item pairs across multiple grades in a K–12 mathematics state accountability test.

A Review of State Assessment Audits: What do States Value? *Michelle Croft, ACT, Inc.; Dan Vitale, ACT, Inc.*

As states seek to implement a balanced assessment system, it is important to understand how they are making decisions about which assessments provide value to stakeholders, including educators, parents, and students. This study examines state assessment audits to identify themes within the audits and actions taken in response to the audit. The results of this study may inform state policymakers and test developers about which aspects of state assessment systems are most valued.

The Impact of Disengaged Test Taking on a State's Accountability Test Results *Sukkeun Im, NWEA; Steven L. Wise, NWEA; Jonghwan Lee, Dr*

This study investigated rapid-guessing behavior on a computer-based state accountability test. Results showed clear evidence of rapid guessing sufficient to meaningfully distort individual student scores and their interpretation. However, both school-level aggregated scores and student proficiency rates were relatively affected by the presence of rapid guessing.

Chair:

Erin Winters, UC Davis

Discussant:

Jon S. Twing, Pearson Assessments

122. Methodological Considerations #2

Paper Session *2:15 to 3:45 pm*

Hilton San Francisco Union Square: Continental 3

Participants:

Composite Creation's Impact on Longitudinal Model's Performance in Detecting Measurement Noninvariance *Dubravka Svetina Valdivia, Indiana University; Rachel Gross, Indiana University; Shimon Sarraf, Indiana University; Jiangqiong Li, Indiana University*

This simulation study examines the impact of composite creation on longitudinal model's performance in detecting longitudinal measurement noninvariance. We found that for many conditions, typically recommended model fit indices (and associated cut off values) failed to detect measurement noninvariance. Implications and future directions are discussed.

Discover Hybrid Connections in Dynamic Network: A Generalized Unified SEM with Regularization Approach *AI YE, UNC Chapel Hill*

Dynamic network can be discovered by fitting Vector Autoregressive (VAR) models on intensive longitudinal data. We extend previous VAR models into a more flexible framework under regularized SEM, that achieves an automatic search for sparse networks with hybrid forms of dynamic relation. Evaluation result from a large-scale simulation is presented.

Estimating Mean and Variance for Stratified Inverse Cluster Samples with Sequential Process *Sewon Kim, Michigan State University*

Stratified inverse cluster sampling (SICS) approach has been proposed to sample rare population. This study compares SICS with sequential process to SICS without sequential process. Both sampling procedures are evaluated depending on sample size and the amount of uncertainty in the population distribution using simulation study.

Evaluating Pretested Item Statistics on Certification Assessments *Sien Deng, ACT Inc.; Han-Wei Chen, ACT, Inc.; Meichu Fan, ACT*

New items are commonly pretested before operational use to get parameter estimates. This study provides a framework to evaluate the stability of item parameter estimates from pretesting to operational testing, and investigates effects of pretested statistics on form assembly and scoring using both empirical data and simulation analysis.

Chair:

Yu Meng, American Osteopathic Association

Discussant:

Tia Fechter, Office of People Analytics

123. Pedagogical Practices

Paper Session *2:15 to 4:15 pm*

Hilton San Francisco Union Square: Franciscan Ballroom A

Participants:

Domain-specificity of Instructional Skills – A Comparative Study in Economics and Mathematics *Christiane Kuhn, Johannes Gutenberg-Universität Mainz; Olga Zlatkin-Troitschanskaia, Johannes Gutenberg-University Mainz; Hannes Saas, Johannes Gutenberg-University Mainz; Jan-Peter Krstev, Johannes Gutenberg-University Mainz*

There is little evidence on the domain-specificity of instructional skills across subjects. Using a newly developed and validated technology-based assessment, the interplay of generic attributes and domain-specific instructional skills of teachers in economics and mathematics is analyzed in a comparative study; and implications for teacher training are drawn.

Measurement Characteristics for Performance Tasks that Assess Content Teaching Skills *Geoffrey Charles Phelps, Educational Testing Service; Brent Bridgeman, ETS*

This session will report pilot results from an administration of 20 newly developed teaching performance tasks with 59 teacher candidates. The session will focus on task latency results, measurement characteristics, participant feedback on the task design, and future use in initial teacher licensure.

Positive Impact of Writing Instruction Using Evidence-Informed Learning Approaches *Heather M. Brown, University of Alberta; Maria Cutumisu, University of Alberta; Chantal Labonté, University of Alberta; Veronica R. Smith, University of Alberta*

Researchers and ten teachers co-developed a persuasive writing unit and examined its impact on middle-school students' writing. The teachers participated in developing and implementing the writing unit and n=292 Grade 5-6 students completed the unit and writing assessments. Findings revealed that the writing intervention improved students' writing skills.

Teachers' Classroom Assessment Practices in the U.S.: Evidence from PISA *Hongli Li, Georgia State University; Roti Chakraborty, Georgia State University*

The PISA teacher questionnaire asked questions about teachers' classroom assessment practices. Drawing on the PISA 2015 U.S. dataset, we addressed two research questions. First, what kinds of classroom assessment practices are used in the U.S.? Second, what teacher-level and school-level characteristics explain the variation of teachers' classroom assessment practices?

The Efficacy of Unfolding Model from a Graded Disagree-Agree Response Scale for Measuring Pedagogical Content Knowledge *Jiwon Nam, Florida State University; Nansook Yu, Chonnam National University; Hyejin Shim, University of Missouri-Columbia*

Despite the importance of PCK, a latent domain on teacher knowledge, it was unknown what IRT model is more appropriate in analyzing PCK response data. This study compared data analyses by IRT models from binary responses. Both 2 PL and unfolding models revealed good model-fit; yet the respondent parameters were differently estimated.

Chair:

Dakota Cintron, University of Connecticut

Discussant:

Stanley N Rabinowitz, Pearson

124. Item Type Considerations

Paper Session *2:45 to 4:15 pm*

Hilton San Francisco Union Square: Franciscan Ballroom C

Participants:

Are All Cognitive Items Equally Prone to Position Effects? *Thai Quang Ong, American Board of Internal Medicine; Dena Pastor, James Madison University*

We examined the relationships among position effects, four item variables, and three person variables simultaneously in two low-stakes assessments. We found easy and long items were most prone to position effects in the low-stakes testing context regardless of examinee gender and effort.

Evaluating Different Scoring Methods for Multiple Response Items Providing Partial Credit *Joe Betts, NCSBN; Doyoung Kim, NCSBN; William Muntean, Pearson; Shu-chuan Kao, Pearson*

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

Modeling Certainty-Based Marking on Multiple-Choice Items: Psychometrics Meets Decision Theory *Qian Wu, KU Leuven (Belgium); Rianne Janssen, KU Leuven (Belgium)*

High-stakes exams administered with certainty-based marking were modeled by the Rasch model (for response accuracy) combined with prospect theory (for the choice of the confidence level and corresponding scoring rule). It is shown that students are not rational decision makers but are influenced by risk attitudes and subjective probability miscalibration.

Predicting Reading Comprehension Item Difficulty Using Cognitive Complexity Features *Maryam Pezeshki, Georgia Institute of Technology; Susan Embretson, Professor*

Understanding sources of cognitive complexity in items is important when designing tests. Previous research indicated strong predictability from cognitive complexity variables in a large-scale reading test. Using the model, cognitive complexity of 400 items from a redesigned test and an older version were predicted. Implications for item design are discussed.

The Comparison Between On- and Off-Grade Items in K-12 Fixed Form and Computerized Adaptive Testing *Deborah Harris, University of Iowa; Qing Xie, ETS*

This study discusses the use of off-grade items in CAT and its comparison with vertically scaled fixed forms. It investigates the impact of item pool characteristics and content constraints on ability estimation, especially for examinees that are of the same ability level but in different grade levels.

Chair:

Priya Kannan, Educational Testing Service

Discussant:

Tracy Lynn Gardner, New Meridian Corporation

113. Electronic Board. Sunday 2:15

Electronic Board Session
2:15 to 3:45 pm
Hilton San Francisco Union Square: Salon A
Participants:

An Investigation of Item Difficulty for Multiple-Texts Reading Skills *Takahiro Terao, The National Center for University Entrance Examinations*

This study aimed to investigate difficulty sources of items to measure reading skills for multiple texts. This study experimentally developed test items with two passages on the same topic. It was shown that, in sentence classification items, a statement describing both two texts was the hardest among all the items.

An Iterative Parametric Bootstrap Approach to Evaluating Rater Fit *Wenjing Guo, University of Alabama; Stefanie Wind, The University of Alabama*

We proposed an iterative parametric bootstrap procedure to overcome limitations of its traditional counterpart for constructing confidence intervals of infit and outfit MSE statistics to identify rater misfit. The proposed procedure was promising because it maintained the false-positive rates at the nominal level and had relatively high true-positive rates.

Benchmarking Early Test-Based Selection in PISA 2009: Lessons for Trinidad and Tobago *Jerome De Lisle, The University of the West Indies*

Using international benchmarking, I compare PISA 2009 data on socioeconomic gradients for 9 comparator countries and Trinidad and Tobago. Different equity effects are seen for early selection and without. However, patterns also vary for systems with test-based selection. Contextual/cultural factors may magnify negative, unintended consequences of early test-based selection

Build Automated Essay Scoring Systems with Advanced Natural Language Processing Techniques *Xiao Luo, Measured Progress*

This study illustrates the process of building an in-house automated essay scoring (AES) system by leveraging recent advances in natural language processing and machine learning. The baseline model indicated a quadratic weighted kappa of .85 in training and .69 in operation. Practical implications and lessons learned will be discussed.

Comparing Methods to Explore Internal Structure: An Examination of Situational Engagement *Daria Gerasimova, George Mason University; Angela Miller, George Mason University; Margret Hjalmarson, George Mason University*

Factor-based methods do not always produce a clear factor solution. Using an example of a multidimensional, situational (instruction-specific) student engagement measure, we examined the ability of non-factor-based methods (Multidimensional Scaling, Item Cluster Analysis, and causal search in TETRAD) to clarify the factor solution, produced by the Exploratory Structural Equation Modeling.

Comparison between Q-function approximation methods in adaptive learning setting *MINGQI HU, University of Illinois at Urbana-Champaign; Jinming Zhang, University of Illinois at Urbana-Champaign*

Adaptive learning system recommends materials according to learners' timely states. Q-function refers to sum of expected rewards from learning process. The study compares two Q-function approximation methods. In simulation, K=2, T=3, two conditions (0, 0.05 measurement error). Fixed-step Q-learning method may outperform Q-function linear approximation method when measurement error occurs.

Comparison of the Performance of Feature Selection Methods in Predicting Bullying Behaviors *Yizhu Gao, University of Alberta; Seyma Nur Yildirim, University of Alberta; Okan Bulut, University of Alberta*

This study compares three commonly used feature selection methods: recursive feature elimination (RFE), principle component analysis (PCA), and correlation-based feature selection (CFS). School crime supplement database was studied. Results show CFS has better performance than PCA and RFE in terms of accuracy (0.84), sensitivity (0.86), and specificity (0.65).

Detecting Neutrality: Critical Consciousness of Educators Scale *Melissa Schneider, University of Denver; Maria Salazar, University of Denver*

Instruments to measure critical consciousness (CC) are scarce and disparate in context and conceptualizations. It is imperative that CC scales detect neutrality toward social identities to identify a continuum of CC. This study uses the "No Opinion" response category to detect neutrality towards social identities in critical consciousness measurement.

Evaluating Effects of Positively and Negatively Worded Questionnaire Items using the Bifactor Model *Walter P. Vispoel, University of Iowa; Guanlan Xu, University of Iowa; Wei S Schneider, University of Iowa; Mingqin Zhang, University of Iowa; Ismail Dilek, University of Iowa; MURAT KILINC, University of Iowa*

We illustrate use of bifactor models with self-report measures in independently quantifying effects of targeted constructs and response styles due to negative and positive phrasing of items. Results highlighted the importance of taking response styles into account when interpreting scores and constructing questionnaires to reduce and balance such effects.

Evaluating Test Form Equivalence of Translated Tests Across Language Groups *Lida Chen, University of Iowa; Hyeon-Joo Oh, Educational Testing Service; Hongwen Guo, Educational Testing Service; Seang-Hwane Joo, Educational Testing Service*

Using the multi-group IRT framework, we examined measurement invariance form equivalence among translated test forms in different languages. Different model fit criteria were used to detect items that may function differently. We also evaluated whether equatings were necessary for different subjects with translated forms.

Examining the Effect of the Number of Common Items in Data Fusion *Katerina M Marcoulides Barbour, University of Minnesota*

Data fusion involves merging datasets sharing some common measures or items, permitting more flexible analyses than when examining the data separately. However, exactly how many common items are needed to effectively integrate multiple datasets has not been determined. This study evaluated the effect of using different numbers of common items.

Exploring the Structure and Technical Adequacy of District-Developed School Climate Survey *Deni Basaraba, Bethel School District #52; Pooja Shivrajo, American Board of Obstetrics & Gynecology*

Due to resource constraints, districts may opt to develop their own surveys to gather data. Concerns abound, however, about the psychometric quality of district-developed surveys. This study will explore the technical adequacy of a district-developed school climate survey, including reliability,

validity, and item-level properties. Implications for locally-developed instruments will be discussed.

Modeling and Reporting Assessment Results: Disentangling Different Aspects of Test Performance *Steffi Pohl, Freie Universität Berlin; Esther Ulitzsch, Freie Universität Berlin; Matthias von Davier, National Board of Medical Examiners*

In order to facilitate a fair comparison and a comprehensive description of test performance, we argue for disentangling different aspects of test performance, that is effective ability, effective speed, and response propensity, and for reporting on a profile of these aspects.

Modeling Conditional Dependence between Response and Response Time in the Testlet Model *Zhaojun Li, The Ohio State University; Dandan Liao, American Institutes for Research; Paul De Boeck, Ohio State University; Frank Rijmen, American Institutes for Research*

A model was proposed for addressing conditional dependence between response accuracy (RA) and response time (RT) in testlet data with RTs only available for testlets but not for individual items. The model has been investigated in a simulation study and used for examining RA-RT dependence for a large-scale assessment.

Modified Exploded Logit Model for Forced-Choice Items in ICCS 2009 *Chia-Wen Chen, Center for Educational Measurement at University of Oslo*

This study developed a modified exploded logit model for the forced-choice item in a large-scale assessment of the International Civic and Citizenship Education Study in 2009. Taiwanese teachers and principals' data were fitted to my model, and the result showed the successful convergence of estimation.

Network DINA Q-matrix Estimation *Xinchu Zhao, Imbellus, INC; Yao Xiong, Imbellus Inc.; Marty McCall, Imbellus Inc.; Jack Buckley, Imbellus, Inc.*

This study proposes a DINA Q-matrix estimation using a network psychometric generalization of the CDM model - the conjunctive root causes model. The estimation is computationally efficient and can be completely exploratory without specifying the number of latent constructs. The fraction subtraction data set is used to demonstrate the estimator.

Test Frequency, Stakes, and Feedback in Student Achievement: A Meta-Regression *Richard P Phelps, Nonpartisan Education Group*

We summarize separate and joint contributions to student achievement via robust variance estimation of three treatments and their interactions with database spanning a century, comprising 149 studies and 509 effect size estimates. Moderators include hundreds of features comprising various test designs and test administration, demographic, and source document characteristics.

The Psychological Network Analysis of Factors Correlating with Physics Ability in NAEQ *Xiaoyue Xiong, Beijing Normal University; Tao Xin, Beijing Normal University, PRC.; Han Du, University of California, Los Angeles*

This research applies a psychological network analysis with Graphical Lasso algorithm to explore the structure among students' physics ability and related factors assessed in NAEQ. The analysis distinguishes the factors that strongly correlate with physics ability (e.g., interest in physics), and the factors that are weakly correlated (e.g., SES).

Using Bayesian Approximate Measurement Invariance Approach to Tracking Student Performance in Mathematics *Shanshan Wang, Fayette County Public Schools*

This methodological case study sought to apply the This methodological case study sought to apply the Bayesian Approximate Measurement Invariance approach to studying 15-year-old students' performance in PISA 2009 and PISA 2012 mathematics tests. The findings of this study demonstrated the flexibility and capacity of the random items effect model to track student performance. to studying 15-year-old students' performance in PISA 2009 and PISA 2012 mathematics tests. The findings of this study demonstrated the flexibility and capacity of the random items effect model to track student performance.

Variational Inference for Cognitive Diagnosis Models *Feng Ji, University of California, Berkeley; Benjamin Deonovic, ACTNext by ACT; Jimmy de la Torre, Hong Kong University; Gunter Maris, ACTNext*

Word Count (47/50) This study proposes the use of variational inference, a fast Bayesian inference alternative to Markov Chain Monte Carlo, for cognitive diagnosis models. The proposed method is comparable to Expectation-Maximization when the number of attributes (K) is moderate and remains computationally feasible when K is large.

Validation of Graph Theory Approach to Detect Test Collusion Network Using an Experimental Real Dataset *Cengiz Zopluoglu, University of Miami; Dmitry Belov, LSAC; James Wollack, University of Wisconsin*

A graph theory approach has been introduced and applied to detect group of test takers who are collectively involved in test collusion (Belov & Wollack, 2018). In this study, we explore the utility of this new approach using real experimental dataset with students known to be engaged in test collusion.

National Council on Measurement - ITEMS Module 3 *Andre Alexander Rupp, Educational Testing Service (ETS)*

Learn about your opportunity to publish in the ITEMS series. The ITEMS portal is your entry point into a world of learning in educational measurement and assessment. ITEMS modules are its centerpiece, which are short self-contained lessons with various supporting resources that facilitate self-guided learning, team-based discussions, as well as professional instruction.

126. **Invited Session Using Educational Assessments to Educate: A Conversation with Edmund Gordon**

Coordinated Session *4:35 to 6:05 pm*

Hilton San Francisco Union Square: Continental 4

If learners vary in their cultural experiences, appreciations, characteristics, and needs, all aspects of culturally relevant pedagogy may need to reflect such variations. However, educational assessments have been most resistant. This session will examine the pros and cons of a repurposing of educational assessment in the service of teaching and learning when learners differ. In this session, Dr. Edmund Gordon will give a brief overview of his current thoughts regarding how educational assessments can be better used to improve education for all students. An invited panel will provide brief reactions to Dr. Gordon's remarks, followed by a facilitated question-and-answer period from the audience.

Presenters:

Edmund Gordon, Teachers College - Columbia University, Yale University
Lloyd Bond, University of North Carolina-Greensboro (Emeritus); Senior Scholar Carnegie Foundation for the Advancement of Teaching (Ret.)
Keena Arbuthnot, Louisiana State University
Richard Duran, University of California, Santa Barbara
Kristen DiCerbo, Pearson

Session Organizer:

Edmund Gordon, Teachers College - Columbia University, Yale University

Chair:

Jennifer Randall, University of Massachusetts Amherst

127. Score Reporting in an Era of Distraction: Elevating Audience and Purpose to Make Measurement Matter

Coordinated Session *4:35 to 6:05 pm*

Hilton San Francisco Union Square: Imperial Ballroom A

No matter how well-designed the assessment, without score reports or interpretation guides that are designed to support the intended purpose and use, we have failed the students, parents, educators, and other stakeholders who we are meant to serve. In this "Research Blitz" panel discussion, we will explore issues related to creating reports for assessments that support accurate, understandable, and actionable inferences and communicating results in ways that best serve the intended purpose and use. The format of the session will be as follows: Each participant will have 5-7 minutes to summarize the particular issue of their choosing related to score reporting. The moderator will then facilitate a discussion among panelists and audience participants. Each participant will bring a different perspective to the table so that, together, the range of assessment types that inform education are represented (formative, interim, summative).

Presenters:

E Caroline Wylie, Educational Testing Service
Kristen Huff, Curriculum Associates
Marianne Perie, Measurement in Practice, LLC
Andrew Ho, Harvard Graduate School of Education

Participants:

How Does the Concept of "Score Reporting" Apply to Formative and Classroom Assessment? *E Caroline Wylie, Educational Testing Service*

Using Interim Assessment Results to Inform Actions in the Classroom *Kristen Huff, Curriculum Associates*

Reporting State Assessment Results: Considering All Stakeholders *Marianne Perie, Measurement in Practice, LLC*

Aggregate Score Reporting: Lessons Learned in Reporting NAEP and SEDA Scores for Schools, Districts, and States *Andrew Ho, Harvard Graduate School of Education*

Session Organizer:

April Zenisky, University of Massachusetts Amherst

Chair:

April Zenisky, University of Massachusetts Amherst

128. The Changing Landscape of Statewide Assessment: Shifts towards Systems of Assessments

Coordinated Session *4:35 to 6:05 pm*

Hilton San Francisco Union Square: Imperial Ballroom B

The landscape of statewide, large-scale educational assessment is shifting away from "stand-alone" summative assessments and towards integrated sets of assessments designed to support various interpretations and uses. For example, several states have provided interim assessments as a part of their statewide assessment program, either individually or as members of a consortium. This coordinated session will explore how the theory of systems of assessments is being applied in multiple contexts, and provide insight into challenges and opportunities inherent in developing and implementing integrated sets of assessments in real world settings. An overview will be provided on developments in theory and practice of balanced systems of assessments (e.g., Pellegrino, Chudowsky & Glaser, 2001), emphasizing implications for current practice. Other presentations focus on ongoing initiatives taking place under the Innovative Assessment Demonstration Authority waivers granted to Louisiana and Georgia. These states aim to replace single statewide summative assessments with multiple assessments that work together to produce a single summative score. This type of assessment model has been referred to as through-course (e.g., Wise, 2011) and might be seen as interim (Dadey & Gong, 2017), but at its core the model is organized around the same principles as balanced systems of assessments.

Presenters:

Brian Gong, Center for Assessment
Abby Javurek, NWEA

Participants:

On the Shift Towards Balanced Assessment Systems: Past, Present and Future *Brian Gong, Center for Assessment*

Developing a Validity Research Agenda for Louisiana's Innovative Assessment Demonstration Authority Pilot *Nathan Dadey, Center for Assessment; Michelle Boyer, Center for Assessment*

On the Opportunities Provided by Through-Year Assessment Models, Including a Solution Configured for Districts in Georgia *Abby Javurek, NWEA; Paul Nichols, NWEA*

Session Organizer:

Nathan Dadey, Center for Assessment

Discussant:

Carla Evans, Center for Assessment

129. Conceptual and Statistical Assessment Linking

Coordinated Session *4:35 to 6:05 pm*

Hilton San Francisco Union Square: Yosemite A

A set of novel empirical methods that are attempting to bridge the divide between statistical linking and equating and provide a conceptual alignment of the measurement constructs for the linked assessments are discussed. Such conceptual approach requires an empirically established and validated learning progression so that assessment items, at the centre of the proposed methods, provide direct information about the status and the trajectory of student learning. The mapping of items from different assessments against such a learning progression provides the conceptual basis for the assessment alignment. These methods thus rely on a more fundamental depiction of the student's understanding, extending the linking beyond mere statistical relationships. The methods presented in this session provide set of solutions for conceptual assessments linking that vary in terms of item mapping processes and resource requirements. The proposed methods offer empirical linking solutions when traditional equating methods are not feasible and where there are educational policy requirements to link and benchmark existing assessments to the external competency standards such as those set by the United Nations Sustainable Development Goal 4 Indicator 4.1.1.

Participants:

Aligning Items to Learning Progressions Using Expert Judgement *Nathan Zoanetti, ACER*

Item Benchmarking Method *Ling Tan, ACER; Nathan Zoanetti, ACER*

Comparative Judgment Linking *Goran Lazendic, ACER; Alejandra Osses, ACER; Ray Adams, The University of Melbourne*

Scale Alignment Through Natural Language Processing Domain Models *Daniel Duckworth, ACER; Ling Tan, ACER*

Session Organizer:

Goran Lazendic, ACER

Chair:

Goran Lazendic, ACER

Discussant:

Catherine A McClellan, Clowder Consulting

130. Fair and Valid Assessment of English Learners with the Most Significant Cognitive Disabilities

Coordinated Session *4:35 to 6:05 pm*

Hilton San Francisco Union Square: Yosemite B

This session addresses the fair and appropriate assessment of English learners (ELs) with the most significant cognitive disabilities, a small yet diverse group of students who undertake triple the work of native English speakers without disabilities (Shyyan & Christensen, 2018); these students must learn academic content while developing proficiency in English, and they may use assistive devices and other communication tools for instruction and assessment. As federally required, States must develop alternate assessments of English language proficiency, and these assessments must be based on standards that include knowledge and skills derived from the four domains of speaking, listening, reading, and writing. For these students, language may manifest differently -- speaking, listening, reading, and writing may not be accessed or demonstrated in ways typical of the general student population. Moreover, published research on this student population is limited. Thus, states face significant challenges in meeting federal requirements and ensuring the fair and valid assessment of these students. Presenters will share recent work relevant to the assessment of ELs with the most significant cognitive disabilities. Policy, heuristics for ensuring fair and valid measures, and research findings that can inform principled approaches to assessment design and development for these students will be discussed.

Participants:

- Assessments of ELs with the Most Significant Cognitive Disabilities: Federal Requirements, Heuristics, and Promising Practices *Deborah Spitz, Office of Elementary and Secondary Education, U.S. Department of Education*
- The Individual Characteristics Questionnaire: Understanding ELs with the Most Significant Cognitive Disabilities in K-12 Settings *Laurene Christensen, WIDA at the Wisconsin Center on Education Research*
- Operationalizing Language Domains for ELs with Significant Cognitive Disabilities: Designing Fair and Valid Measures *Edynn Sato, Sato Education Consulting LLC*
- Developing Item Templates for Alternate Assessments of English Language Proficiency *Phoebe Winter, Independent Consultant*

Session Organizer:

Edynn Sato, Sato Education Consulting LLC

Discussant:

Christopher J. Rivera, East Carolina University

131. Using Process Data for Advancing the Practice and Science of Educational Measurement

Coordinated Session *4:35 to 6:05 pm*

Hilton San Francisco Union Square: Yosemite C

The move from paper-based to digitally-based assessments is creating new data sources that allow us to think differently about the foundational aspects of measurement, including sources of evidence for reliability, validity, fairness, and generalizability. Process data, also referred to as “observable data,” or “action data” reflect individuals’ behaviors while completing an assessment task. They are logs of individuals’ actions, such as key strokes, time spent on tasks, and eye movements. These data reflect student engagement with assessment and can provide important insights about students’ response processes that may not be captured in their final solutions to the assessment tasks. The four presentations focus on use of process data (1) in measurement modeling and psychometric models; (2) for enhancing group comparisons and fairness research; (3) to examine differences in the processes proficient and less proficient students use to write essays; and (4) to assess computational thinking captured during gameplay.

Participants:

- Implications of Considering Response Process Data for Psychometrics *Roy Levy, Arizona State Univerisity*
- Use of Response Process Data to Inform Group Comparisons and Fairness Research *Kadriye Ercikan, ETS/UBC; Hongwen Guo, Educational Testing Service; Qiwei He, Educational Testing Service*
- How do Proficient and Less-Proficient Students Differ in their Composition Processes? *Randy E Bennett, ETS; Mo Zhang, ETS; Paul Deane, ETS; Pater van Rijn, ETS*

Session Organizer:

Kadriye Ercikan, ETS/UBC

Chair:

Joan Herman, CRESST/UCLA

Discussants:

James Pellegrino, University of Illinois, Chicago
Joan Herman, CRESST/UCLA

132. Psychometric Considerations in the Measurement of Social-emotional Learning and School Climate

Coordinated Session *4:35 to 6:05 pm*

Hilton San Francisco Union Square: Continental 1

While measures of social-emotional learning (SEL) and school climate are growing in prominence as components of schools' accountability and improvement systems, much less is known about the psychometric quality of the instruments used in practice. In particular, understanding the development of students' perceptions of SEL and climate requires modeling and interpreting growth trajectories, yet little is known about how much common problems with Likert-based measures affect estimates of growth. In this panel, we examine the intersection between self-report issues and a desire to estimate developmental trajectories in two ways. First, we examine the assumptions required to scale SEL and school climate measures for growth inferences, including the appropriateness of various item response theory models to produce scale scores for use in growth models, longitudinal measurement invariance assumptions, and whether the stability of scores is due in part to consistent response styles. Second, we consider the promise and limitations of alternatives to self-report Likert-type items for measuring SEL, in particular using open-ended item responses and text-mining methodologies. Altogether, this session can be useful to measurement experts, practitioners, and policymakers alike by illuminating how much faulty assumptions about self-report measures might affect growth estimates and briefly considering design- and psychometric-based alternatives.

Participants:

Measuring Growth in Students' Social-emotional Learning: A Comparison of Multiple Scoring Approaches *Megan R. Kuhfeld, NWEA; James Soland, NWEA*

Multilevel Longitudinal Measurement Invariance and School Climate Surveys *Jon Schweig, RAND*

Accounting for Students' Socially Desirable Responding in the Measurement of Social-emotional Skills *James Soland, NWEA; Megan R. Kuhfeld, NWEA*

Using Text-Mining Models to Quantify Creative Thinking *Denis Dumas, University of Denver; Peter Organisciak, University of Denver*

Session Organizer:

Megan R. Kuhfeld, NWEA

Chair:

Megan R. Kuhfeld, NWEA

Discussant:

Dan Bolt, UW-Madison School of Education

133. Research Blitz - Teachers and Assessments

Research Blitz Session *4:35 to 6:05 pm*

Hilton San Francisco Union Square: Continental 2

Participants:

A measurement approach to harvest actionable information from tests to improve teaching *Lucy Fang Lu, NSW Department of Education, Australia; Wai-Yin Wan, Department of Education, NSW, Australia*

In response to challenges faced by teachers in getting actionable information from online adaptive test data, this paper proposes a generalised DIF approach to harvest instructionally relevant information from the test data to help inform teaching. Modelling results are presented to demonstrate the utility, reliability and validity of this approach.

Assessing Dimensionality of a High-Stakes Teacher Candidate Performance Assessment: poly-DETECT and CFA *Aileen Reid, University of North Carolina at Greensboro; Lexi Lay, University of North Carolina at Greensboro; Kristen Smith, University of North Carolina at Greensboro*

This study compared poly-DETECT and Confirmatory Factor Analysis procedures to examine the factor structure of a high-stakes teacher candidate performance assessment. Results from both methodologies provided initial support for the same measurement model. Utilizing results from both methods allowed researchers to make more informed decisions about the instrument's factor structure.

Assessment Fluidity: Helping Teachers to View Externally Developed Assessments as Instructional Tools *Valeria Carolina Zunino, University of California, Davis; Megan E Welsh, University of California, Davis*

Using qualitative data from Chile's first large scale interim assessment we present evidence showing that teachers classify assessments into two excluding groups: large-scale assessments for accountability purposes and classroom assessments for formative purposes. We discuss how this conceptualization may inhibit the use of externally developed assessments designed for classroom use.

Assessment of teacher practices and student achievement *Anh Hua, Rutgers University; Adam Lekwa, Rutgers University; Linda Reddy, Rutgers University*

We examine how teachers' teaching strategies are associated with elementary school students' reading growth and whether the same strategies are equally effective across grades. Preliminary results show that teachers' classroom strategies are associated with students' reading performance within each year and that the strategies are more influential among older students.

Developing Test Performance Communication Solutions in a Teacher-Researcher Partnership *Chad M Gotch, Washington State University; Mary Roduta Roberts, University of Alberta*

This collaborative study aimed to develop solutions for teachers to communicate to families about test performances. We will share teacher-developed tools to support this task, lessons learned, next steps in the partnership, and reflections on our approach to engaging teachers as equal partners.

Exploring the Relationships among Teacher Feedback, Self-Regulation, and Academic Achievement *Yongfei Wu, Queen's University*

This study found that students' conceptions of teacher feedback (SCoTF) were significantly correlated with their self-regulation and classroom performance, but not with other academic achievement indicators. SCoTF significantly but negatively predicted self-perceived English level. Students' self-regulation positively predicted their classroom assessment results but not with their external standardized test scores.

Measuring Pedagogical Content Knowledge (PCK) of High School Math Teachers (HSMT) *Tadeu Aparecido Pereira da Ponte, Insper; Jorge Herbert Soares de Lira, Universidade Federal do Ceará; Thomaz Edson Veloso, Universidade Federal do Ceará; Francisco Bruno de Lima Holanda, Universidade Federal de Goiás; Angelica Turaça, Insper*

Measuring teacher's PCK for effective math instruction had advanced in elementary grades, mostly with multiple choice items tests. This proposal summarizes the efforts to develop measures for PCK of HSMT, based on multiple types of items, in order to define and pursue teacher readiness for in service teacher's development programs.

Online Surveys for Measuring Teachers' Assessment Literacy and Conceptions of Assessment *Kim Koh, University of Calgary*

Existing measures of teachers' assessment literacy and conceptions of assessment are insufficiently sensitive to inform teachers' professional learning needs and pedagogical practices. This paper reports on our development and validation of two online survey instruments to measure teachers' assessment literacy and conceptions of assessment for diagnostic purposes.

Relationship between Teachers' Informal Assessment, Teaching Strategies and Student Reading Achievement *Lu Guo, Texas Tech University*

While much attention has been given to the (in)effectiveness of informal assessments, there has been relatively less research on whether teachers' beliefs regarding informal assessment are aligned with their teaching instructions. This study specifically examines the relationship among teachers' instructions, informal assessments and students' reading performance.

Robust Estimation of Teacher Effects with Multiple Raters: Use of t-distributional Assumptions *Michael Seltzer, UCLA; Jayashri Srinivasan, UCLA; Minjeong Jeon, UCLA*

Multiple raters are employed in many educational settings. We illustrate a fully Bayesian cross-classified modeling approach that employs t-distributional assumptions, which will downweight a discrepant rating received by a teacher vis-à-vis other ratings received by the teacher, thus providing some robustness in drawing inferences concerning teachers' level of instructional quality.

Using Performance Tasks to Assess Content Teaching Skills *Mellissa Fowler, ETS; Geoffrey Charles Phelps, Educational Testing Service*

An ECD approach was used to identify and assess the knowledge and skill used in teaching --e.g., evaluating student work, representing content, modeling concepts. We will share an example of a newly developed performance task and related pilot results from an administration of 20 tasks with 59 teacher candidates.

Chair:

Brian C Leventhal, James Madison University

134. Research Blitz - Differential Item Functioning

Research Blitz Session *4:35 to 6:05 pm*

Hilton San Francisco Union Square: Continental 3

Participants:

An Application of Differential Item Functioning Methods: Across Time Comparison *Ruben Castaneda, College Board*

We explored possible construct irrelevant variance in a large-scale English language proficiency examination administered across various institutions in Latin America. Three years of data were evaluated for differential item functioning (DIF) using a step-up IRT model fit approach. Results and implications of DIF testing in English language will be discussed.

Diagnostic Assessment: DIF Detection from Binary and Aggregate Mastery Classifications *Jeffrey Hoover, University of Kansas; William Thompson, University of Kansas; Amy K Clark, University of Kansas; Brooke Nash, University of Kansas*

This study examines differential item functioning (DIF) procedures and matching variables impact DIF detection for diagnostic assessment systems. Items from a large-scale, K-12 assessment were analyzed using logistic regression and Mantel-Haenszel methods. The Mantel-Haenszel procedure identifies more items with non-negligible DIF with less consistency than the logistic regression procedure.

Differential Item Functioning for Polytomous Response Items Using Hierarchical Generalized Linear Model *Meng Hua, University of Pittsburgh; Feifei Ye, RAND Corporation*

This study applied hierarchical generalized linear model (HGLM) as a DIF assessment method to polytomous response items. Results were compared to the Generalized Mantel-Haensel (GMH) procedure and logistic regression. HGLM showed promising advantages in selecting DIF-free items as anchor items and controlling for the Type I error of DIF detection.

Differential item functioning with Multistage Testing *Ru Lu, Educational Testing Service; Paul A Jewsbury, ETS*

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

Ecological Framework for Item Response of a Youth Risk and Needs Assessment *Thao Vo, Washington State University; Brian F French, Washington State University*

An ecological differential item functioning (DIF) framework investigated a risk and needs assessment using contextual administrative school data to explain DIF variance across three ethnic groups. Utilizing two multilevel models, significant family and school contextual effects were found. Results demonstrate how context helps explain DIF beyond traditional grouping categories.

Evaluation Anchor Items on DIF using Longitudinal TIMSS Data (2007 – 2015) *Youn-Jeng Choi, University of Alabama; Wenjing Guo, University of Alabama*

The purpose of this study is to explore what may be contributing to differences in performance of anchor items in mathematics and science on the TIMSS datasets during 2007 - 2015. This will be done by using a mixture IRT modeling approach to first detect latent classes in the data and then to examine differences in performance on items taken by examinees in the different latent classes.

SIBTEST Cut-Scores: Reevaluating the Classification Guidelines for Dichotomous DIF *James D Weese, University of Arkansas*

The utilization of SIBTEST for differential item functioning (DIF) has been used for over 25 years. The current cut-scores associated were derived from a limited simulation study, and they potentially are erroneous. The aim of this paper is to investigate the cut-scores with a more robust, and inclusive simulation study.

Testing Measurement Invariance using Multi-Group SEM: Equivalence Testing and a Projection-Based Method *Ge Jiang, UIUC*

Multi-Group Structural Equation Modeling is commonly utilized to examine measurement invariance. However, a logical issue arises when nonsignificant chi-square statistics are used to confirm measurement invariance. I introduce an equivalence testing approach to examine measurement invariance and a projection method to compare latent means when scalar invariance fails to hold.

The Impact of Source of Matching Abilities on DIF in Multistage Testing *Sakine Gocer Sahin, WIDA at WCER at UW-Madison; Kyoungwon Bishop, WIDA at WCER at UW-Madison; Halil I Sari, Kilis 7 Aralik University; Metin Bulus, Adiyaman University*

The purpose of this research is to investigate the source of measure of ability to match students in MST design. Specifically, we examine whether the matching abilities of reference and focal groups estimated from a final stage or from previous stages impact DIF or not.

Transadaptation socio-emotional skills assessments to geographical regions: Can It be done? *Cristina Anguiano-Carrasco, ACT, Inc.; Jason Way, ACT, Inc*

Given the increasing interest in social and emotional skills assessments globally, researchers and practitioners are faced with the challenge of transadaptation. The process is essential to ensure fairness, but it demands many resources. Transadapting to multiple contexts that are culturally close can result on important savings.

Using Propensity Score Methods in Determining the Causes of Differential Item Functioning *Kubra Atalay Kabasakal, NO; Terry Ackerman, University of Iowa*

The aim of this study is to determine the causes of DIF by using the propensity scores of equivalent groups in non-experimental research. In this study, we used PISA international educational test data which is frequently used in DIF studies and contains many heterogeneous subgroups.

Chair:

Bozhidar M. Bashkov, American Board of Internal Medicine

135. Research Blitz - Student Assessment

Research Blitz Session *4:35 to 6:05 pm*

Hilton San Francisco Union Square: Franciscan Ballroom A

Participants:

A Framework for Evaluating a Learner At-Risk System *Eryka Nosal, Kaplan Professional; Anna Topczewski, Kaplan Professional*

Stakeholders must interpret course results and decide who is at-risk of failing and when to intervene. The study purpose is to compare at-risk models in terms of model accuracy, interpretability, and ability to provide remediation. This study can be used as a framework for such an evaluation.

Algebra Ready or Not: Establishing and Validating a Cut Score Along a Learning Progression *Jennifer Lee Lewis, University of Massachusetts Amherst*

This study investigated the effectiveness of two methods for identifying a cut score along a learning progression. The reliability and validity of each cut score are presented as evidence for the selection of the best method to determine a cut score along a learning progression within a formative assessment context.

Assessing Socio-affective and Cognitive Dimensions of Learning: Exploring a Strategy of Data-Based Decision-Making *Sandra Zepeda, Universidad Catolica de Chile; Valeria Carolina Zunino, University of California, Davis*

A comprehensive assessment of socio-emotional and cognitive dimensions was administered to all the students of a school and a data analysis plan was designed and implemented. The results show that the use of information through socialization and literacy programs in the educational community can favor the decision-making process.

Coding structured interviews to place students along a learning progression. *Amy Elizabeth Cardace, Cornell University; Mark Wilson, University of California, Berkeley; Kathleen E. Metz, University of California, Berkeley*

This paper describes the mapping of students' interview responses to a learning progression in elementary science. We explore the complexities of the coding process and present in-depth inter-rater analyses to highlight their importance. Our findings also provide evidence that the process yielded codes that were ultimately reliable and valid.

Development of a Learning Progression to Support the Assessment of Intercultural Capability *Rebekah Luo, The University of Melbourne; Toshiko Kamei, The University of Melbourne; Masa Pavlovic, The University of Melbourne; Bruce Beswick, The University of Melbourne; Eeqbal Hassim, The University of Melbourne*

The research described in this proposal developed an assessment based on a learning progression of intercultural capabilities. There is a dearth of resources targeting intercultural capabilities due to the difficulty of measuring and defining such skills. Thus, this study applied a curriculum and instruction driven approach to address such challenges.

Early Detection of High School Dropout: A Deep Learning Approach with Keras *Okan Bulut, University of Alberta; Soo Lee, American Institutes for Research*

Increasing high school dropout rates have been an important issue for school systems. Advanced predictive models can be utilized for predicting high school performance and understanding factors leading to the dropout issue. This study employs a deep learning approach with Keras to predict dropout status using a longitudinal dataset.

Examining three psychometric models for evaluating learning progressions: Simulation and empirical perspectives *Duy Ngoc Pham, Educational Testing Service; Craig S. Wells, University of Massachusetts Amherst*

This study aimed to examine a multi-dimensional IRT model and two cognitive diagnostic models for evaluating learning progressions using simulated and empirical data. Simulation results indicated that the model complemented each other. The empirical analysis provided convergent evidence to support almost all the aspects of the theory underlying two progressions.

Exploring Learning Pathways in a Critical Online Reasoning Assessment *Olga Zlatkin-Troitschanskaia, Johannes Gutenberg-University Mainz; Susanne Schmidt, Johannes Gutenberg University Mainz; Dimitri Molerov, Humboldt Universität zu Berlin; Marie-Theres Nagel, Research Assistant*

Within the computer-based Critical Online Reasoning Assessment undergraduate students are asked to follow one URL (e.g. a twitter link) and to judge whether the information is reliable by conducting a free web search. Written statements from the participants, along with event data, enables us to visualize and evaluate learning pathways.

Keeping It on the Level: Using Learning Trajectories for Diagnostic Inferences *Sanford Student, University of Colorado Boulder; Derek Briggs, University of Colorado Boulder; Emily Toutkoushian, North Carolina State University; Jere Confrey, North Carolina State University*

This study investigates the validity and reliability of a mathematics assessment with items designed to align to two learning trajectories. We examine the extent to which empirical item difficulties align with learning trajectory levels and the impact of this assessment's length on data visualizations, a key basis for classroom decision-making.

Making Early Warning Systems Whole: Evaluating Prediction Models Using Social-Emotional Learning Data *David Alexandro, Connecticut State Department of Education*

The purpose of this study is to determine the impact of including social-emotional learning (SEL) variables and other cross-disciplinary predictors on the classification accuracy of models identifying which students are at risk of not being proficient in reading, and to identify which variables predict reading proficiency in third grade.

Scoping Literature on Interactions and Intersections of Learning Assessment and Student Motivation *Kai Jun Chew, Virginia Tech*

This study scopes the literature on interactions and intersections of learning assessment and student motivation in engineering education, though findings can inform similar reviews on the topic in the larger education context. Preliminary findings show small size and extent of such literature in engineering education.

The Influencing Mechanism of Instructional Quality on Students' Mathematical Learning Behavior *LU YUAN, Beijing Normal University; XIAOFENG DU, Beijing Normal University; XUEFENG LUO, Minnan Normal University; Tao Xin, Beijing Normal University, PRC.*

This research uses SEM to investigate the influence mechanism of instructional quality on students' mathematical learning behavior based on the data

PISA 2012 for Shanghai, China. The results showed that mathematical interest played a mediating role in the relation between instructional quality and students' mathematical learning behavior.

Chair:

Matthew Gaertner, WestEd

136. Research Blitz - Computer Based Testing

Research Blitz Session *4:35 to 6:05 pm*

Hilton San Francisco Union Square: Franciscan Ballroom C

Participants:

A New Approach to a Pretesting Design in Multistage Testing *Yanming Jiang, Educational Testing Service*

We propose to conduct pretesting at both stages of a two-stage multistage test (MST). Simulation results show that the proposed methods have advantages over the commonly used Stage 1-only pretesting method in terms of estimation accuracy of pretest item parameters and examinee abilities.

Administration of Selective Domain Tests in Benchmark Adaptive Testing *Rong Jin, Riverside Insights; Unhee Ju, Riverside Insights; JP Kim, Riverside Insights.com*

The mathematics domains in a grade generally differ in difficulties and time of the academic year being taught to students. This study explores the effects of selective domain tests on ability estimates in CAT by comparing to the original full test.

Developing Field Test Designs using Targeted Multistage Techniques *Mingqin Zhang, University of Iowa; John Denbleyker, HMH; Shuqin Tao, American Institutes for Research*

The study proposes a two-step field test plan and investigates four calibration designs: linear, multistage, targeted linear, and targeted multistage designs. The goal is to evaluate each calibration design in terms of calibration efficiency and cost-benefit of implementation. A simulation study is conducted based on a real operational item bank.

Effect of Different Blueprint Constraint Levels on an Adaptive Test *Jonghwan Lee, Dr; Sukkeun Im, NWEA; Christina Schneider, NWEA*

Adaptive tests must adhere to the test blueprints and be able to recover students' true ability. The test engine administers items based on various constraints that ensure coverage of the blueprints. A recent study evaluated results from two constraint levels to investigate the impact each had on the adaptive administration.

Evaluating Psychometric Properties of Computerized Multistage Testing *Shumin Jing, University of Iowa; Won-Chan Lee, University of Iowa*

The main purpose of this study is to propose formal procedures of estimating psychometric properties for computerized multistate testing (MST), including conditional standard errors of measurement, reliability, and classification consistency and accuracy. A comprehensive evaluation of these properties under various conditions is also presented.

Feasibility of Transitioning to MST with a Bank Designed for Fixed Forms *Tracy Lynn Gardner, New Meridian Corporation; Richard Luecht, University of North Carolina; Leslie Keng, National Center for the Improvement of Educational Assessment*

This purpose of this study was to investigate the feasibility of multistage testing (MST) for an item bank. MST offers the potential for enhanced score precision across score scales. This study simulated the outcomes of building MST test forms given a variety of content constraints and other test assembly specifications.

Field Test Item Calibration Under Multistage Testing *Sophie Cai, Cognia; Louis Roussos, Measured Progress*

In this study, we propose a field test item calibration procedure that intends to maximize the preservation of newly-developed items under a 1-3 MST structure.

Impact Automated Scoring Engine Upgrade on Scoring *Zhen (Jane) Wang, Educational Testing Service; Gautam Puhan, ETS*

Due to the automated scoring engine upgrade, there is a need to check the impact. We will also check if the engine upgrade is substantial such that it warrants an equating adjustment to put the scores obtained using the new and old scoring engines on the same reporting scale.

Investigating Students' Calculator Use and Computations Using Process Data in Mathematics Assessment *Manqian Liao, University of Maryland, College Park; Fusun Sahin, American Institutes for Research*

This study empirically examines how students use the calculator to solve mathematics items in digitally-based NAEP. Computations students performed with the calculator were extracted from process data and compared with expert anticipations. Results could inform calculator access decisions in assessments and provide diagnostic information about computational strategies.

Precision control of CAT item exposure through enemy relationships *Kirk Alan Becker, Pearson; Haiqin Chen, American Dental Association; QIAO LIN, University of Illinois at Chicago*

Overuse of items in a CAT pool is a concern for both security and content representation. This paper will demonstrate the application of enemy item exclusions to exposure control. This approach can be easily applied in addition to any other exposure control methods, and can target specific items.

The Impact of the First Passage Selection in Passage-based Reading CAT *David Shin, Pearson; Jadie Kong, Pearson*

The purpose of this study is to investigate the impact of the first passage selection in Reading CAT tests. There are two first passage selection options investigated. Preliminary results show that using all available first passage candidates yield similar test precision but better item exposure rate and passage usage.

Using Out-of-Level Items in Formative ELA CAT: Benefits and Constraints for High and Low Performers *Jie Lin, Pearson*

This study uses item pools of different sizes (both in-level and out-of-level) to investigate the benefits and constraints of using out-of-level items for high and low performers in formative ELA CAT. Results have direct implications with regards to the relevance and applicability of using out-of-level items in ELA CAT.

Chair:

Jonathan Rubright, NBME

125. GSIC - Electronic Board. Sunday 4:35

Graduate Student Issues Committee (GSIC)
Graduate Electronic Board Session
4:35 to 6:05 pm
Hilton San Francisco Union Square: Salon A
Participants:

An Analysis of TIMMS Data with Diagnostic Classification Models *Selay Zor, active; Laine Bradshaw, University of Georgia*

Large-scale assessments are valuable sources to obtain information about education on national/international level. Subscores used to describe performance provide limited-information(Sinharay&Haberman,2010).To investigate the degree to which more detailed-information may be garnered from large-scale assessments, we reanalyze an existing large-scale assessment data using DCMs to provide latent variable-based diagnostic information on examinees' cognitive attributes.

Application of the Hybrid Model to Determine Utility of Process Data *Clifford Erhardt Hauenstein, Georgia Institute of Technology; Matthew Johnson, Educational Testing Service; Jie Gao, Educational Testing Service*

Action sequences from the 2015 NAEP Science Assessment are used to gather information regarding examinee problem solving ability. However, action sequences may not represent valid indicators of response process equally well for all examinees. The utility of the Hybrid model (Yamamoto, 1989) in clustering examinees accordingly is proposed and applied.

A Simulation Study on IRT-Based Vertical Equating Methods *Yan Yan, Graduate Student; Susan Embretson, Professor*

This study examines the performance of several unidimensional and multidimensional item response theory models on the vertical linking of simulated item response data with a longitudinal design. It is hypothesized that a multidimensional item response theory model for change will yield the most plausible results for equating.

Automated Essay Scoring with Multi-Level-Representation Deep Models *Chang Lu, University of Alberta; Mark J. Gierl, University of Alberta*

This study implemented three deep learning algorithms with different levels of representations for prompt-dependent automated essay scoring tasks. Results for the comparisons among the three models show that Convolutional Neural Network + Long Short-Term Memory model with word and sentence representation layers demonstrate highest accuracy and efficiency.

Comparing the Efficiency of DIF Detection for HGLM, MIMIC, and MNLFA *Yue Yin, University of South Florida; Eunsook Kim, University of South Florida*

We explored the efficiency of DIF detection for HGLM, MIMIC model and MNLFA model in a categorical and a continuous covariate under different conditions. The Monte Carlo simulation was conducted for generating a total of 528 conditions. Result found HGLM exhibited higher power in detecting categorical uniform DIF.

Detecting Guessing Behaviors by Response Time in Computerized Multistage Testing *YIBO WANG, University of Iowa; Mingjia Ma, University of Iowa; Deborah Harris, University of Iowa; Stephen Dunbar, University of Iowa*

Response time provides us another way to detect aberrant behaviors in the test. MST collects each examinee's RTs for each item during the test. This study investigated using multistage testing design under different levels of guessing and speed and determines the effectiveness of using response time to detect guessing behaviors.

Estimating Reliability of Literacy Assessments: A Multivariate Generalizability Theory Approach *Kuo-Feng Chang, University of Iowa; Jaime Malatesta, Center for Advanced Studies in Measurement and Assessment; Won-Chan Lee, University of Iowa; Robert Brennan, Professor Emeritus, University of Iowa, Center for Advanced Studies in Measurement and Assessment; Po-Hsi Chen, National Taiwan Normal University*

Estimating reliability for literacy assessments can be difficult due to the use of multiple types of questions and item formats (e.g., multiple-choice, true-false, constructed-response, and Likert-type items). Therefore, this study demonstrates the potential advantages of using multivariate generalizability theory to estimate reliability for a literacy assessment with complex item types.

Evaluating Four Approaches to Handling Zero-Frequency Scores under Equipercentile Equating *Ting Sun, University of North Carolina at Charlotte; Stella Yun Kim, UNC Charlotte*

The purpose of this study is to compare four methods in equipercentile equating involving zero-frequency scores: adding a minimal relative frequency, using a middle score, using log-linear pre-smoothing, and borrowing frequencies from its adjacent score points. A simulation study is conducted to evaluate these methods in terms of equating accuracy.

Examining Relation of Information and Communication Technology to Collaborative Problem Solving by Hierarchical Linear Modeling *Yizhu Gao, University of Alberta; Fu Chen, University of Alberta; Ying Cui, University of Alberta*

This study examined relationship between information and communication technology (ICT) and collaborative problem solving (CPS) performance. The Programme for International Student Assessment (PISA) 2015 database was used. Results show 'schoolwork usage' negatively predicts students' CPS scores; 'game usage' also negatively affected CPS performance; however, 'communication usage' positively affect CPS score.

Exploring the Moderation Effects of Technology on Gender DIF of Reading Comprehension *Yue Yin, University of South Florida; Yi-Hsin Chen, University of South Florida; Robert Dedrick, University of South Florida*

We explored gender differential item functioning (DIF) of 7 English speaking countries in reading comprehension test for investigating the moderation effects of technology and DIF sources in multilevel structure. Result found gender DIF items varied across countries and technology had moderation effect for gender DIF.

Fine Tuning Factor Analyses to Improve Model Fit at the Item Level *Wei S Schneider, University of Iowa; Ismail Dilek, University of Iowa; Walter P. Vispoel, University of Iowa; MURAT KILINC, University of Iowa; Guanlan Xu, University of Iowa; Mingqin Zhang, University of Iowa*

Confirmatory factor analyses often fail to yield satisfactory model fits due to unwarranted assumptions of equal-interval measurement, failure to account for method factor variance, and use of overly restrictive factor models. We demonstrate how these issues can be addressed using data obtained from popular self-concept, personality, and social desirability scales.

Investigating the Skill Hierarchies in Reading Comprehension *Xuejun Ryan Ji, The University of British Columbia; Amery D Wu, The University of British Columbia; Kausalai Wijekumar, Texas A&M University College Station*

This study investigated the skill hierarchies in reading comprehension by comparing DCMs. Various dependence structures among the three skills of literal, inferential, and critical, borne on levels of comprehension theory, were tested. Preliminary results indicated that literal comprehension was the prerequisite skill for both inferential and critical comprehension.

Looking Validity in Cognitive Diagnostic Model (CDM) through Differential Item Functioning (DIF) *Gamze Kartal, University of Illinois at Urbana-Champaign; Jinming Zhang, University of Illinois at Urbana-Champaign*

To check the validity and test fairness of a test, a DIF analysis should be conducted. However, there is a shortage of research on DIF in CDM context. This study aims at comparing DIF detection methods using DINA and generalized DINA models when DIF presents.

Missing Data and The Rasch Model: The Effects of Missing Data Mechanisms on Item Parameter Estimation *Tom Waterbury, James Madison University*

This study explores the effects of missing data on Rasch model item parameter estimates. When responses were missing completely at random or missing at random, item parameters were unbiased. When responses were missing not at random, item parameters were severely biased, especially when the proportion of missing responses was high.

Practical Applications for Detecting Item Preknowledge and Compromised Content Through Sequential Monitoring *Linette P Ross, student*

This research will use data forensics to evaluate three statistical methods using responses and response times to detect compromised content for continuous monitoring of items in an operational exam with frequent test administrations. This research will leverage information about known security breaches, testing patterns, and well-defined subgroups to strengthen investigations.

Predictive Analytics of Temporal Behavior to Automate Formative Feedback on Course Performance *Fu Chen, University of Alberta; Ying Cui, University of Alberta*

In this study, students' temporal behaviors in the learning management system were modeled with long short-term memory (LSTM) networks to automate immediate feedback on course performance. Results showed that compared with the prediction by aggregated behaviors with conventional machine learning classifiers, the LSTM approach demonstrated higher accuracy and stronger generalizability.

Response Time Variation in PISA 2015 Science *Emily Kerzabi, Educational Testing Service*

Using PISA 2015 Science data, this research simultaneously investigates response time (RT) patterns by item type, item difficulty, and cognitive demand while controlling for participant ability and response accuracy (correct/incorrect). Such RT patterns should be considered when setting RT thresholds, conducting quality checks, or assessing RT models.

Testing Access to the Test? Perceptions of Standardized Tests by Adults with Dyslexia *Maura O'Riordan, University of Massachusetts Amherst*

This paper explores the experiences that adults with dyslexia recall having with standardized tests. The purpose of this paper is to identify ways to improve the testing process for these students to ensure validity in test scores for students with dyslexia and improve current practices.

The Effect of Repeat Item Exposure in a High-stakes Standardized Assessment *Xiaodan Tang, University of Illinois at Chicago; Matthew Schultz, AICPA*

This study aims to examine the potential impacts on repeat examinees' performance by reusing performance assessment items in a high-stake standardized assessment. We found that there are limited benefits from encountering the same items. We also discussed the practical implications to the licensing or certification assessments.

The Impact of Undesirable Distractors on Estimates of Ability *Kathryn N Thompson, James Madison University; Brian C Leventhal, James Madison University*

Little attention has been placed on how the type of distractor influences ability estimates. Polytomous IRT models provide test-makers with a way to examine distractor functioning. We use simulation to evaluate the effects from three types of undesirable distractors by examining bias in ability estimates.

The NIDA-Power Model for Iterated Attributes *Xiaoliang Zhou, Columbia University; Yi Chen, Teachers College, Columbia University; James Corter, Columbia University*

Iterated attributes in a test item can increase item complexity, hence difficulty. We propose the NIDA-Power model, where the application probability of an attribute for an item is determined by both the attribute's slip parameter and the number of times the attribute is used. Simulations show reasonable performance.

The School Performance Framework: What is Really Being Measured? *Kaitlin Mork, University of Colorado Boulder*

School Performance Framework (SPF) ratings are calculated using a complex scoring scheme involving disaggregation and weighting. Analyses are conducted to investigate (1) whether SPF ratings can be accurately predicted using a simpler, ordered logit model and (2) whether demographic variables are predictive of SPF ratings above and beyond academic indicators.

Using Auxiliary Item Information in the Item Parameter Estimation of a Graded Response Model *Matthew David Naveiras, Vanderbilt University; Sun-Joo University Cho, Vanderbilt University*

This paper applied empirical and hierarchical Bayes methods using auxiliary item information to a GRM to obtain item parameter estimates with greater stability and precision than MMLE, particularly in small to medium sample sizes. The effectiveness of these Bayes methods relative to MMLE was evaluated via a simulation study.

Using Rapid Responding to Evaluate Test Speededness *Yage Guo, University of Nebraska-Lincoln*

This study is aimed at using rapid responding to evaluate test speededness on high-stakes testing by investigating examinees' response behaviors. The study will also investigate examinee and item characteristics related to rapid responding. The process data are obtained from a high-stakes licensure test, including diverse examinee's response behaviors during test.

National Council on Measurement - ITEMS Module GSIC 1 *Andre Alexander Rupp, Educational Testing Service (ETS)*

Learn about the opportunity to publish in the ITEMS Module series. The ITEMS portal is your entry point into a world of learning in educational measurement and assessment. ITEMS modules are its centerpiece, which are short self-contained lessons with various supporting resources that facilitate self-guided learning, team-based discussions, as well as professional instruction.

SUNDAY APRIL 19

139. Indicators of Educational Equity: Tracking Disparities, Advancing Equity

Coordinated Session *8:15 to 10:15 am*

Hilton San Francisco Union Square: Continental 4

In 2017, the National Academies of Sciences, Engineering, and Medicine appointed a committee to identify key indicators for measuring the extent of equity in the nation's K-12 education system. The committee proposed 16 indicators classified into two categories: measures of disparities in students' academic achievement, engagement, and educational attainment; and measures of equitable access to critical educational resources and opportunities. The committee recommended ways to develop and implement a system to measure and monitor equity over time. They called for indicators to be collected on a broad scale across the country with reporting mechanisms designed to regularly and systematically inform stakeholders at the national, state, and local levels about the status of educational equity in the United States. The system they envisioned would rise to the level of priority given to NAEP, involving both government and non-governmental entities in its development. The primary objectives for this session are to disseminate the findings from this study and stimulate interest in doing the work needed to develop and implement a system of equity indicators. This session will consist of three formal presentations and one that is interactive with audience members.

Participants:

Presentation 1: Rationale for the Study *Christopher J. Edley, UC Berkeley*

Presentation 2: Indicators of Disparities in Student Outcomes *Laura Hamilton, RAND*

Presentation 3: Indicators of Inequitable Access to Educational Resources *Sean F. Reardon, Stanford University*

Presentation 4: Dissemination, Implementation *Natalie Nielsen, National Academies of Sciences, Engineering, and Medicine*

Session Organizer:

Judith A Koenig, National Academies of Sciences, Engineering, and Medicine

Discussant:

Andrew Ho, Harvard Graduate School of Education

140. Artificial Intelligence in Educational Measurement: Trends, Mindsets, and Practices

Coordinated Session

8:15 to 10:15 am

Hilton San Francisco Union Square: Imperial Ballroom A

In this session, various experts from areas connected to artificial intelligence (AI) in assessment will provide thoughtful perspectives on how key issues in educational measurement are conceptually framed, empirically investigated, and critically communicated to key stakeholder groups. After a general overview of the current trends in AI research as it pertains to educational assessment, different presenters will critically discuss how the three core areas of (1) reliability / statistical modeling, (2) validity / construct representation, and (3) equity / fairness are tackled in a changing field of educational assessment. A related goal of the session is to have presenters and participants suggest key lines of work to which current members of NCME can productively contribute to shape best practices in this new world of assessment. In addition, the session will be used as an opportunity to discuss means of cross-community outreach and engagement that can help build further bridges between the current NCME membership and members from neighboring scientific and practitioner communities working with AI technologies in assessment. This session is connected to a newly proposed SIG "Artificial Intelligence in Assessment".

Participants:

AI in STEM Assessment: Trends, Mindsets, and Practices *Janice Gobert, Apprendis; Mike Sao Pedro, Apprendis*

AI in Education: New Data Sources and Modeling Opportunities *Piotr Mitros, ETS; Steven Tang, eMetric*

Fairness, Accountability, and Transparency in Machine Learning *Collin Lynch, North Carolina State University*

Bias and Fairness for Automated Feedback Generation *Neil Heffernan, Worcester Polytechnic Institute; Anthony Botelho, Worcester Polytechnic Institute*

Building and Picking a Model for Learning and Assessment *Michael Yudelson, ACTNext by ACT*

Session Organizers:

Andre Alexander Rupp, Educational Testing Service (ETS)
Carol M Forsyth, Educational Testing Service

Discussants:

Alina von Davier, ACTNext
Andre Alexander Rupp, Educational Testing Service (ETS)
Carol M Forsyth, Educational Testing Service

141. **Invited Session Fireside Chat with Classroom Assessment Task Force: Making Measurement Meaningful with Classroom Assessment**

Coordinated Session – Panel Discussion *8:15 to 10:15 am*

Hilton San Francisco Union Square: Imperial Ballroom B

In 2016, NCME President Mark Wilson established a Classroom Assessment Task Force (CATF) "to support and facilitate the integration of classroom assessment into the NCME consciousness and scholarship" (CATF, 2017). Specifically, the vision of the CATF was articulated in the five-year charter submitted to the NCME Board in 2017 as: NCME membership and scholarship (conference presentations, training workshops, journals) will reflect a balance of research and theory on the full range of uses of educational assessment. NCME seeks to influence classroom assessment practices through the appropriate application of measurement principles and insights and for these principles and insights to be influenced by classroom practitioners (CATF, 2017). The session is designed to facilitate dialogue among members of the Task Force, leaders in the field of Classroom Assessment, and NCME members. Details on the format of the session are described below. The intent of the session is to examine the degree to which the CATF has made progress towards the articulated vision and to identify ways to close potential gaps between that vision and our current activities.

Presenters:

Sue Brookhart, Dusquene University
Heidi Andrade, SUNY Albany
Mark Wilson, University of California, Berkeley
E Caroline Wylie, Educational Testing Service
Jade Caines Lee, University of New Hampshire
Alison Bailey, UCLA
Neal Kingston, University of Kansas
Dale Whittington, Retired

Session Organizer:

Kristen L Huff, Curriculum Associates

142. Use of Natural Language Processing to Support Tutoring

Coordinated Session *8:15 to 10:15 am*

Hilton San Francisco Union Square: Yosemite A

Educational measurement plays an important role in learning and navigation systems by seeking who you are, where you are at, and where you are going. The use of computers and mobile devices makes it possible to track students' progress and diagnose their mastery in real time and render feedback for personalized guidance and intervention. As technology advances, natural language has become increasingly important in educational measurement. It is prevalent in student-tutor online and offline conversations, as well as in open-ended responses to different types of assessments. Unlike selected responses in yes-or-no or multiple-choice questions, natural language not only provides an intuitive representation of one's understanding, but also serves as a natural tool for communicating comprehensible feedback and information. With improvements in virtual assistants in particular that only engage in audio conversations, there is a growing interest in creating an automated tutoring system which interacts with students via personalized and adaptive dialogue grounded with a shared representation. In this symposium, we present and discuss related research and development in diagnosing student's misconception in learning and generating adaptive guidance in forms of a dialogue, hint, or video, based on natural language processing. We will provide practical solutions to real-world problems in education.

Participants:

Machine Learning for Scalability and Improvement of Tutoring Systems *Jay Thomas, ACT, Inc.; Sue Ward, ACT; Meirav Arrieli-Attali, ACT*

Mining Misconceptions from Natural Language Logs in Tutoring Systems *Zachary Pardos, UC Berkeley*

A Dialogue Dataset for Multi-Conversation Foreign Language Tutoring Interactions *Katherine Stasaski, UC Berkeley; Marti Hearst, UC Berkeley*

Generation of Micro Multimodal Content *Yuchi Huang, ACT; Saad Khan, ACTNext by ACT*

Session Organizer:

Lu Ou, ACT

143. Exploring Scoring and Psychometric Modeling Strategies for Innovative Items

Coordinated Session *8:15 to 10:15 am*

Hilton San Francisco Union Square: Yosemite B

As assessments have transitioned from paper-and-pencil formats to digital formats, testing programs have been increasingly including innovative or technology enhanced (TE) items for a variety of purposes. These items can be classified across several dimensions of innovation, including: item format, response action, media inclusion, level of interactivity, and scoring method (Parshall, Davey, & Pashley, 2000). This session explores various ways in which the first four dimensions may influence the scoring method dimension, defined as the method by which examinee responses are translated into quantitative scores (Parshall et al., 2000). The scoring dimension incorporates both scoring rules to determine the number and order of score levels for innovative items and the modeling strategy employed. Each of the papers in this session includes analysis of real data, which together encompass a wide variety of innovative item types from several testing programs. The session includes commentary from a leading researcher in psychometric methods.

Participants:

A Framework for Rule-Based Scoring of Technology Enhanced Items *William Lorie, Center for Assessment*

Exploring Strategies for Optimal Scoring Rubric Development of Technology Enhanced Items *Adrienne Sgammato, Educational Testing Service*

Item Response Theory Models for Adaptive Testlet Items *Carol Eckerly, Educational Testing Service; Paul A Jewsbury, ETS; Yue Jia, Educational Testing Service*

A Comparison of Rapid Picture and Rapid Color Naming Screeners *Adam E Wyse, Renaissance; Scott McConnell, University of Minnesota, Twin Cities; Eric Stickney, Renaissance Learning; Catherine N. Close, Renaissance; Heidi Lund, Renaissance*

Evaluating Bias from Introducing New Items into a Scale *Paul A Jewsbury, ETS; Ru Lu, Educational Testing Service*

Session Organizer:

Carol Eckerly, Educational Testing Service

Discussant:

Billy Skorupski, Amira Learning

144. Principled Item Design: State-of-the-Art

Coordinated Session *8:15 to 10:15 am*

Hilton San Francisco Union Square: Yosemite C

This session provides an overview of some of the state-of-the-art methods and operational as well as validity challenges associated with implementing principled item design procedures using cognitive modeling of item difficulty, artificial intelligence/machine learning, and automatic item generation. A variety of testing settings will be covered with empirical examples and research results provided—including assessments used in achievement, placement and cognitive reasoning. The session will allow participants time to ask questions of the presenters.

Participants:

Test and Item Development with Cognitive Design Systems: Some Examples *Susan Embretson, Professor*

The interplay of validity and efficiency *Isaac I Bejar, Educational Testing Service*

Considerations for training subject matter experts to write item models *Audra Kosh, Edmentum*

The challenges of principled item design *Richard Luecht, University of North Carolina*

Automated item generation using deep learning *Matthias von Davier, National Board of Medical Examiners*

Session Organizers:

Richard Luecht, University of North Carolina
Isaac I Bejar, Educational Testing Service

Chair:

Richard Luecht, University of North Carolina

Discussant:

Steve Ferrara, Measured Progress

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

Coordinated Session *8:15 to 10:15 am*

Hilton San Francisco Union Square: Continental 1

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

Participants:

Joint Modeling of Response Times and Item Response Data using a GLM Approach *Daniella Alves Rebouças, University of Notre Dame; Ying Cheng, University of Notre Dame*

Empirical Investigation of Joint Modeling of Responses and Response Times in Linear-on-the-Fly Tests with Testlets *Hong Jiao, University of Maryland; Xin Qiao, University of Maryland, College Park; Jung-Jung Lee, University of Maryland, College Park*

A Mixture Response Time Process Model to Detect Aberrant Behaviors and Explain Item Nonresponses *Jing Lu, Northeast Normal University; Chun Wang, University of Washington*

Utilizing Response Time to Measure Person Slipping in High-Stakes Tests *Yang Du, University of Illinois, Urbana-Champaign; Justin L. Kern, University of Illinois at Urbana-Champaign*

A Machine Learning Approach to Modeling Response Times *yiqin pan, University of Wisconsin-Madison; Edison M Choe, Graduate Management Admission Council*

Session Organizer:

Huijuan Meng, Graduate Management Admission Council

Discussant:

Edison M Choe, Graduate Management Admission Council

146. Test Equating

Paper Session *8:15 to 10:15 am*

Hilton San Francisco Union Square: Continental 2

Participants:

Bifactor IRT True-Score Equating Under Nonequivalent Groups Design *Kyung Yong Kim, University of North Carolina at Greensboro; Uk Hyun Cho, University of North Carolina at Greensboro*

Under the common-item nonequivalent groups design, this study presents a procedure for equating passage-based tests with the bifactor model based on projective item response theory. The feasibility of the proposed procedure is assessed through simulation under various study conditions, including degrees of local item dependence and distribution of latent variables.

Evaluating Several Variations of Simple-Structure MIRT Equating *Stella Yun Kim, UNC Charlotte; Won-Chan Lee, University of Iowa*

The primary purposes of this study are to propose a new observed-score equating method under simple-structure multidimensional IRT (SS-MIRT) and compare it with other variations of SS-MIRT equating methods. Both observed-score and true-score equating methods are considered and evaluated with respect to equating accuracy under several simulation conditions.

Feasibility of Using International Test Takers for Test Equating *Tom Waterbury, James Madison University; Ying Lu, College Board; Judit Antal, College Board*

In this study, we explored the effects of using international test-takers as one of the equating groups for equating forms of a high-stakes assessment. Results indicated that equating relationships were accurate only in certain conditions, namely, when using a mini/semi-midi anchor set with a very large number of anchor items.

Is pre-smoothing necessary in kernel equating? *Tom Benton, Cambridge Assessment*

This study explores whether pre-smoothing using log-linear models is a necessary precursor to equating using the kernel method of test equating, or whether a suitable choice of bandwidth within the method can provide sufficient smoothing. The accuracy of the different approaches was compared using data from 515 real assessments.

The effect of Differential Item Functioning and Anchor-Total Relationships for Testlet-Based Equating *Nixi Wang, University of Washington*

Equating has been widely used in large-scale assessments, where testlet items are common for test design. To ensure the validity of the results from equating methods, differential item functioning (DIF) and anchor-total relationships of items are investigated in a study of testlet equating design. Two different R packages were used to simulate data and conduct equating through separate and IRT equating methods. Results in terms of equating error (RMSE) and bias (BIAS) are presented and complexities of multidimensional testlet IRT equating are discussed.

Chair:

Tzur Karelitz, National Institute for Testing & Evaluation (NITE)

Discussant:

Jonathan Weeks, ETS

147. Emerging Psychometric Models and Methods

Paper Session *8:15 to 10:15 am*

Hilton San Francisco Union Square: Continental 3

Participants:

Correcting Bias in G-Theory Variance-Component Estimates When Additivity Assumption Is Violated *CHIH-KAI LIN, American Institutes for Research; Jinming Zhang, University of Illinois at Urbana-Champaign*

Estimation precision of variance components in G theory is of great importance because they serve as the foundation of estimating generalizability/reliability coefficients. The current study compared five different methods of correcting bias in variance-component estimates in the presence of varying degrees of nonadditivity.

Exploring the "Cluster-Independent" Property of Ability Estimators in Bifactor Models *Dandan Liao, American Institutes for Research; Frank Rijmen, American Institutes for Research; Tao Jiang, American Institutes for Research*

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

The Effect of Non-normal Latent Distributions on Model Selection Methods *Tzu-Chun Kuo, American Institutes for Research; Yanyan Sheng, Southern Illinois University Carbondale*

Seven model selection indices were compared in fitting unidimensional data with normal or non-normal latent traits. Preliminary results suggested that skew affects more on their performances than kurtosis, and that BIC is preferred without guessing whereas WAIC and LOO are preferred when guessing is involved in item responses.

The Effect of Trend Rescore Design on New Monitoring Statistics *John Donoghue, Educational Testing Service*

In trend scoring, a set of Time A responses are rescored by Time B raters. Constraints on the resulting table violate assumptions underlying, e.g., the paired t-test. This simulation study examined whether using responses already rescored (within Time A) as the trend sample could improve detection of rater drift.

The Influences of Ability on Response Accuracy: Using the Full Cross-loadings Hierarchical Model *Maoxin Zhang, University of Oslo*

A full cross-loadings hierarchical model (FCHM) was proposed to concurrently model response times and response accuracy (RA), by adding cross-loadings to the hierarchical framework. A real data study and a simulation study indicated FCHM recovered some item and person parameters better and showed the best goodness-of-fit.

Chair:

Kimberly Colvin, University at Albany, SUNY

Discussant:

Ken Fujimoto, Loyola University Chicago

148. Large-Scale Assessment Programs

Paper Session *8:15 to 10:15 am*

Hilton San Francisco Union Square: Franciscan Ballroom A

Participants:

Can Examinees Adequately Perceive Barriers to Performing Well? Device Comparability Insights *Tia Fechter, Office of People Analytics; Daniel O Segall, Office of People Analytics*

A device comparability study was conducted for ASVAB administration. Following a counterbalanced random assignment of two device conditions to examinees, post-test feedback was collected to determine if the device negatively impacted performance on a subset of ASVAB questions. Analysis will determine whether examinees adequately report feedback that correlates with performance.

Evaluating Effects of Extended Time Accommodations for NAEP: Multiple Control Groups *Youmi Suk, Department of Educational Psychology, University of Wisconsin-Madison; Peter M. Steiner, Department of Educational Psychology, University of Wisconsin-Madison; Jee-Seon Kim, Department of Educational Psychology, University of Wisconsin-Madison*

This study examines the effects of extended time accommodations (ETA) on math proficiency using NAEP Grade-8 Mathematics 2017 assessment data. We define the effects of ETA using causal diagrams and suggest conditional instrumental variables and propensity score matching to estimate the causal effects.

Exploring Students' Perspective in a Digitally Based Assessment Using Process Data *Fusun Sahin, American Institutes for Research; Juanita Hicks, American Institutes for Research; Shuang (Grace) Ji, American Institutes for Research*

Self-reported test-taking experiences were examined in digitally-based NAEP Assessment between students who received two different types of test forms. Comparisons were made between students who received items adapted from the paper-based version and those who received items designed for digitally-based assessment by evaluating process data, survey responses, and scores together.

Imputations in Large-Scale Assessment Contextual Data: Is Recreation of Plausible Values Necessary? *Ting Zhang, AIR; Paul Bailey, AIR; Sinan Yavuz, Mr.; Huade Huo, AIR*

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

Writing Performance and Digital Familiarity: Multi-Group SEM Approach *R. Noah Padgett, Baylor University; Young Yee Kim, American Institutes for Research; Xiaying (James) Zheng, American Institutes for Research; XIAOYING FENG, Avar Consulting, Inc; American Institutes for Research (contractor)*

Using multi-group structural equation modeling, this study explores if digital familiarity measured by prior exposure to writing on a computer is related to writing performance and if the relationship varies across major subgroups (sex and race). The NAEP 2011 grade 8 writing digitally based assessment data were used.

Chair:

Jaime Malatesta, Center for Advanced Studies in Measurement and Assessment

Discussant:

Maria Elena Oliveri, Educational Testing Service

149. Test Design Considerations

Paper Session *8:15 to 10:15 am*

Hilton San Francisco Union Square: Franciscan Ballroom C

Participants:

Choose your Pictures Wisely: The Role of Representational and Decorative Pictures in Educational Assessment *Marlit Annalena Lindner, IPN - Leibniz Institute for Science and Mathematics Education*

How do visualizations, such as representational or decorative pictures affect students' item processing, performance and motivation in educational assessment? Two experiments that employ computer-based assessment and eye tracking with over five-hundred children, give insights into students' interaction with different picture types during test-taking. Implications for item design are discussed.

Comparisons Among Approaches to Match Non-Equivalent Equating Samples *Sooyeon Kim, ETS; Michael Walker, College Board*

This study compares three approaches for reducing group nonequivalence in a situation where not only is randomization unsuccessful but the number of common items is limited. Group adjustment through either subgroup weighting, a weak anchor, or a mix of both, is evaluated in terms of equating accuracy.

Cross-classified Random Effects Modeling for Moderated Item Calibration *Seungwon Chung, University of Minnesota, Twin Cities; Li Cai, University of California—Los Angeles*

Test forms are often modified to accommodate various special populations or situations where administration of the original test forms is infeasible. As a systematic method for obtaining comparable scores across test forms, this study proposes a unified cross-classified random effects model to revise item parameters for scoring modified test forms.

Determining blueprint classifications using value added subscores *Chen Qiu, University of Kentucky; Michael Peabody, American Board of Family Medicine; Rongxiu Wu, University of Kentucky*

When designing tests, subject matter experts often prescribe a blueprint classification system without fully considering whether the categories provide useful feedback to examinees. This study examines several possible blueprint classification systems with differing category structures and examined each to determine the best method for communicating useful subtest feedback to examinees.

Practical Implications of Quasi-Experimental Design Choice *Edgar I Sanchez, ACT*

This research examines the practical implications of using propensity score, coarsened exact, and Mahalanobis distance matching (MDM) and inverse probability of treatment weighting to estimate the effects of using a test preparation workbook.

Chair:

Anna Topczewski, Kaplan Professional

Discussant:

Anne Traynor, Purdue University

138. Electronic Board. Monday 8:15

Electronic Board Session *8:15 to 10:15 am*

Hilton San Francisco Union Square: Salon A

Participants:

Are Students Fairly and Equally Assessed in Science Education? *Haiying Li, ACT, Inc.*

Inequity in science education is a pressing issue. This study focused on inequity from an assessment perspective. Results showed approximately half of students were unfairly and unequally assessed when the assessment adopted either selected-response or constructed-response formats. Implications for how to enhance equity in science education were discussed.

Assessing group collaboration in small teams *Nafisa Awwal, The University of Melbourne; Mark Wilson, University of California, Berkeley; Patrick Griffin, The University of Melbourne*

The literature shows measures of group collaboration have not been explored adequately. The authors propose group collaboration measures from a process perspective of collaborative problem solving. Generic indicators including autoscoring algorithms are developed to measure group collaboration in small teams. Empirical data will examine the validity evidence for these measures.

California Superintendents' Beliefs About School Climate Assessment for Continuous Improvement: Multi-dimensionality and Response Patterns *Yidan Zhang, UC Berkeley; Anji Bucker, San Jose State University; Brent Duckor, San Jose State University; Mark Wilson, University of California, Berkeley*

This study measures three distinct constructs of superintendents' beliefs regarding school climate data (Importance, Capacity, and Trustworthiness). IRT analysis confirmed the multidimensionality of the scale and revealed the patterns of superintendents' responses across these three belief constructs. providing a foundation for improving assessment practice through understanding and influencing leaders' beliefs.

Comparing norming methods performance under violation of the assumptions *Ou Zhang, Pearson; Xuechun Zhou, NCS Pearson; M. Kuzey Bilir, Pearson; Xiaolin Wang, NBOME*

We examined which norming method provide precise and acceptable norm estimates, under different conditions of particular sample size, violations of normality and homoscedasticity. the four norming methods. The quality of the norms was estimated with multiple indexes such as polynomial curve fit, overall statistics, IPR, score distributions. Norming method performance differences will be illustrated under different conditions (i.e., sample size, violation of normality or homoscedasticity) to establish a solid common understanding of selecting norming methods.

Dealing with Bifactor-Structured Data: Effect Of Sample Size On IRT Model Selection *XINYUE LI, Penn State University*

Researchers are faced with model selection problem between the simpler models and complex models, especially when sample size is limited. This study is to examine whether parameter estimates produced by the simpler unidimensional model would be comparable to those produced by bifactor model for testlet data under different sample sizes.

Designing and Evaluating a Diagnostic Classification-based Concept Inventory in Middle Grades Statistics *Laine Bradshaw, University of Georgia; Madeline Schellman, University of Georgia; Jessica Masters, Research Matters, LLC; Lisa Famularo, Research Matters, LLC; Hollylynne Lee, North Carolina State University; Hamid Sanei, North Carolina State University*

Through the IES-funded Diagnostic Inventories of Cognition in Education (DICE) project, our interdisciplinary team designed a concept inventory of probabilistic reasoning that leverages diagnostic measurement techniques for maximizing the inventory's efficiency and accuracy for diagnosing misconceptions. We report on validity evidence gathered via 60 cognitive interviews and a large-scale administration.

Evaluating Conditional False Positive Rates of Answer-Copying Indices *Hui Deng, College Board*

This study evaluates conditional false positive rates for several copy-detection indices, using examinee pairs simulated at different ability levels based on data from a large scale Reading assessment. The results have implications for the choice and implementation of copy-detection indices in applied settings.

Exploring DIF for EL Students with Disabilities in an English Proficiency Test *Kyoungwon Bishop, WIDA at WCER at UW-Madison; Sakine Gocer Sahin, WIDA at WCER at UW-Madison; Hyeon-Joo Oh, Educational Testing Service*

An increasing number of students in the EL population have disabilities. This study examines performance differences between English Learner (EL) students with disabilities and EL students without disabilities in differential item function (DIF) in a large scale ELP assessment, with an eye toward fairness in assessment.

Maintaining Score Scales: A Comparison Study *Won-Chan Lee, University of Iowa; Stella Yun Kim, UNC Charlotte*

This study explores various scoring methods including number-correct scoring, IRT pattern scoring, and hybrid scoring in relation to their capability of maintaining scale scores over time. A simulation study is conducted to evaluate relative performance of six scoring methods in reproducing expected scale-score moments and passing rates.

Measuring change for a longitudinal assessment *Fen Fan, NCCPA; Joshua Goodman, National Commission on Certification of Physician Assistants; Drew Dallas, NCCPA*

The goal of this study is to examine the effectiveness of a longitudinal assessment piloted by a medical certification board. Specifically, this study will explore if examinees' performance change over time and what variables (e.g., specialty, practice setting, and years of experience) can help interpret the performance change.

Measuring Interdisciplinary Research Skills and Opportunity: A Construct Modeling Approach *Cheryl Schwab, University of California, Berkeley*

We investigated the impact of a graduate research training on participants' perceived ability and opportunity to do interdisciplinary research at the intersection of global environmental change and data science. The results of a survey created through a construct modeling approach show the relationship between interdisciplinary research skills and with opportunity.

Screening for aberrant school performances in high-stakes assessments using influential analysis *Andrés Christiansen, KU Leuven*

A method is proposed for screening aberrant school performances in large-scale, high-stakes assessments. Proportions of high achievers within a school were modeled via the beta inflated mean regression model and a measure of f-divergence to determine aberrancy. A simulation study revealed that the method can recover distorted school performances.

Short Test Form Assembly Using Aggregated Classification Models with Class Imbalance Remedies *Xuechun Zhou, NCS Pearson*

The purpose of this study is to investigate the effectiveness of using aggregated classification models for short test form assembly when severe class imbalance is present. Three classification tree models with three post hoc sampling remedies and one cost-sensitive training are built with the objective to maximize minority class predication.

Strategies for Implementing CD-CAT in High-Dimensional Testing Situations *Yan Sun, Rutgers University; Jimmy de la Torre, Hong Kong University*

CD-CAT has been developed to administer diagnostic tests more efficiently; however, when the number of attributes is large (i.e., test is high-dimensional), implementing CD-CAT becomes infeasible. To address the high-dimensionality issue, a strategy that involves item calibration, modified item selection and shrinking the prior distributions is proposed.

The Effect of School Switch in Chile: A Cross-Classified Modeling Approach *Luo Jinwen, UCLA; Maria Paz Fernandez, UCLA*

This study proposes a cross-classified random coefficient model to estimate the school switch effects on students' outcomes and differentiate the effects among various switching behaviors. The model has been applied to a Chilean administrative testing data and we found a significant variance of the switch effects across Chilean schools.

The Effects of Including an Unnecessary Specific Factor in the Bifactor Model *Mina Lee, University of Massachusetts Amherst; Scott Monroe, University of Massachusetts Amherst*

This study investigates the effects of specifying a specific factor in the bifactor model when the specific factor does not exist. Based on the simulation results, irregular factor loadings, including both negative and positive directions, occurred when the regular bifactor model was fitted to data without a specific factor.

The Relationship between Growth Percentiles and Growth Projections Overtime *Jessalyn Smith, DRC; Scott Li, DRC*

Growth measures are used in accountability systems. Results classify student growth and/or predict how much the student needs to grow reach or maintain a "proficiency". However, there is a gap in the literature studying the relationship and consistency of growth measures and projections. This study looks to fill that gap.

Two-Level Matching: Can Adding School-Level Characteristics Improve Matching Sampling for Comparability Studies? *Xin Li, ACT, Inc.; Meichu Fan, ACT*

When the implementation of random assignment is impossible, matching sampling can be used as an approximation to the randomization. This study evaluates whether adding school-level characteristics can improve matching quality or not under classical test theory and item response theory, using matched samples comparability analyses and the propensity score matching.

Using the SRMR to determine sample size in structural equation modeling *Alberto Maydeu-Olivares, University of Barcelona*

We discuss how to compute the power of the Standardized Root Mean Residual (SRMR) to reject a closely fitting model when fit is acceptable. The formulae may be used to determine the sample size. We compare our approach to the current standard based on the RMSEA.

Response Styles Analysis with different approaches *Sohee Kim, Oklahoma State University; Hyejin Shim, University of Missouri-Columbia; Ki Cole, Oklahoma State University*

Many research studies are conducted using various surveys; however, there were a few studies about systematic error such as response style biases. Therefore, this study explores response styles in order to avoid biases in interpretations of results applying three different approaches: mixed Rasch, multidimensional partial credit, and IR tree model.

151. ** GSIC Featured Session ** Advice on Transitioning from Graduate Training to the Professional World

Coordinated Session – Panel Discussion *10:35 to 12:05 pm*

Hilton San Francisco Union Square: Continental 4

In this featured panel organized by the Graduate Student Issues Committee (GSIC), panelists across different measurement industry sectors will discuss their experiences transitioning from graduate students to working psychometricians. Panelists will share advice they wish they had known as graduate students. Additionally, they will be asked questions about the skills and knowledge important for successful professionals in the measurement field. A dedicated a question and answer portion will be included in the session.

Presenters:

Juan D'Brot, NCIEA
Kadriye Ercikan, ETS/UBC
Susan Lottridge, AIR
Matthew Madison, Clemson University
Melinda Taylor, ACT

Session Organizer:

Maura O'Riordan, University of Massachusetts Amherst

Chair:

Delwin Carter, University of California, Santa Barbara

152. Making Assessment Matter: Improving the Assessment Literacy of State and District Policy Leaders

Coordinated Session *10:35 to 12:05 pm*

Hilton San Francisco Union Square: Imperial Ballroom A

Assessment literacy efforts have focused on improving the knowledge and skills of educators to better design and use assessments to improve instruction for students. Unfortunately, there has been little work focused on improving the assessment literacy among state policy makers, yet much of the blame for assessment system incoherence arguably falls on state and district leaders—the decision-makers regarding assessment choices. The implementation of balanced assessment systems requires that leaders understand the features of high-quality, balanced assessment systems at all levels: classroom, district, and state. Unfortunately, policy makers lack critical skills about how assessments should fit together as a system to support multiple users and uses in a coherent and efficient manner. Given the powerful influence of state assessment laws and regulations on the design and implementation of large-scale assessment systems, it is surprising how little we, as a measurement community, know about these laws and regulations. This interactive, coordinated session brings together three related papers and two leaders in our field to serve as discussants to further the conference theme of making measurement matter by focusing on how to improve assessment literacy among those who make assessment policy.

Participants:

Assessment Policy Analysis *Zachary Feldberg, University of Georgia*

Assessment Literacy for Policy Makers *Brittney Hernandez, University of Connecticut*

Policy levers to support balanced assessment systems *Scott Marion, National Center for the Improvement of Educational Assessment*

Session Organizer:

Scott Marion, National Center for the Improvement of Educational Assessment

Discussants:

Lorrie Shepard, University of Colorado
Andrew Middlestead, Michigan Department of Education

153. What is the Value Proposition for Principled Assessment Design?

Coordinated Session *10:35 to 12:05 pm*

Hilton San Francisco Union Square: Imperial Ballroom B

Principled assessment design approaches (PAD), such as evidence centered design, have been available to the field of educational assessment for almost two decades. The use of PAD appears to offer many benefits including improved validity evidence, more efficient assessment development, and support for innovative assessment approaches. Yet, PAD does not dominate operational assessment design and development in educational assessment. Pieces are implemented here and there but no operational program has implemented all the elements of PAD. In this session, the adoption of PAD is hypothesized to be driven by the value created for customers. The lack of PAD adoption suggests that customers perceive little value in using PAD. The presenters will explore the value proposition for PAD, where a value proposition is the value or need that PAD is fulfilling for customers. The first presenter will describe the components of a PAD business model including the users and buyers for PAD, the problems for which they might use PAD to help address, and the channels for communicating with these customers. The following three presenters will propose who they view as the customer, describes the key drivers of value, identify blockers to use, and suggest a means to improve PAD adoption.

Participants:

Business Model for Principled Assessment Design *Paul Nichols, NWEA*

The Only Job to be Done is Helping Teachers Teach and Students Learn *Kristen Huff, Curriculum Associates*

No Need to Tear Down When You Can Build Up *Steve Ferrara, Measured Progress*

PAD or not to Pad: Let the Market Decide *Catherine Needham, NWEA; Christina Schneider, NWEA*

Session Organizer:

Paul Nichols, NWEA

Discussants:

Jeremy Heneger, Nebraska Department of Education
Rhonda True, Nebraska Department of Education

154. Assessing Collaborative Problem Solving at Scale: The Status Quo and the Next

Coordinated Session *10:35 to 12:05 pm*

Hilton San Francisco Union Square: Yosemite A

Collaborative problem solving (CPS) is widely accepted as an essential 21st-century skill and is crucial for success in academia and the workplace. However, there is a lack of generally available assessment instruments for CPS to support the teaching and learning of CPS such that CPS can be accurately defined, and improvements can be quantitatively measured. Developing a large-scale and standardized assessment of CPS that can be administered regularly is extremely challenging. Over the last decade, leveraging advances of computer and internet technology, several significant efforts, such as the Assessment and Teaching of 21st Century Skills (ATC21S) and the 2015 Programme for International Student Assessment (PISA 2015), have been made to assess CPS at scale. Meanwhile, educational testing companies, such as ETS and ACT, also launched their extensive research agenda to explore CPS assessments. This coordinated session includes five presentations from researchers who have been deeply involved in the efforts as mentioned above, as well as from the computer-supported collaborative learning community, to present the status quo in CPS assessment. We hope these updates can help the community to understand where we are now, what are the main challenges, and what our next steps should be to achieve synergistic advances.

Participants:

CPS assessment in ATC21S - learning from the past *Esther Care, Brookings Institution*

Updates on CPS Assessment in PISA 2015 *Art Graesser, University of Memphis*

Assessing CPS: Conceptualization and Methodological Differences between the CSCL and Assessment Communities *Nancy Law, University of Hongkong*

Learning & Assessment of Collaborative Problem Solving at ACT *Alina von Davier, ACTNext; Kristin Stoeffler, ACTNext by ACT; Benjamin Deonovic, ACTNext by ACT; Michael Yudelson, ACTNext by ACT; Pravin Chopade, ACTNext by ACT; David Edwards, ACTNext by ACT; Saad Khan, ACTNext by ACT; Yigal Rosen, ACTNext by ACT*

ETS' Efforts towards Learning and Assessments of CPS at Scale *Jiangang Hao, Educational Testing Service*

Session Organizer:

Jiangang Hao, Educational Testing Service

Discussant:

Patrick Kyllonen, Educational Testing Service

155. A Case Study in Measurement Practice and the Public Perception

Coordinated Session *10:35 to 12:05 pm*

Hilton San Francisco Union Square: Yosemite B

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

Participants:

Validity - What We Do *Emily Shaw, College Board*

Fairness – What We Do *Michael Walker, College Board*

Accountability – What We Do *Walter (Denny) Way, College Board*

Session Organizer:

Emily Shaw, College Board

Discussants:

Brent Bridgeman, ETS
Rebecca Zwick, Educational Testing Service
Ellen Forte, edCount, LLC

156. Modeling Measurement Invariance and Response Biases in International Large-Scale Assessments

Coordinated Session *10:35 to 12:05 pm*

Hilton San Francisco Union Square: Yosemite C

The main goal of international large-scale assessments (ILSAs) – such as PISA, PIAAC, TIMSS, PIRLS – is to provide unbiased and comparable test scores and data which enable valid and meaningful inferences about a variety of educational systems and societies. In contrast to national surveys, ILSAs can provide a frame of reference to extend our understanding of national educational systems and cross-country variability. To enable fair group comparisons (within and across countries) and valid interpretations of statistical results in low-stakes assessments such as ILSAs, two validity aspects need to be accounted for. First, the data need to be tested and corrected for response biases such as response styles (RS) in non-cognitive scales. Second, the comparability of the data and test scores across different countries and languages needs to be established. This is achieved by testing and modelling measurement invariance (MI) assumptions. The proposed coordinated session provides an overview of state of the art and new psychometric approaches to test MI assumptions, handle the problem of response biases, and investigate the relations and interactions between both. The goal is to provide researchers, practitioners and policy makers with comparable and meaningful data for secondary analyses and to enable fair comparisons of groups and countries.

Participants:

A comparison of Multigroup-CFA and IRT-based item fit for measurement invariance testing *Janine Buchholz, Leibniz Institute for Research and Information in Education (DIPF); Johannes Hartig, DIPF | Leibniz Institute for Research and Information in Education, Frankfurt, Germany*

Comparing three-level GLMMs and multiple-group IRT models to detect group DIF *Carmen Köhler, Leibniz Institute for Research and Information in Education (DIPF); Lale Khorramdel, National Board of Medical Examiners; Johannes Hartig, DIPF | Leibniz Institute for Research and Information in Education, Frankfurt, Germany*

Comparability and Dimensionality of Response Time in PISA *Emily Kerzabi, Educational Testing Service; Hyo Jeong Shin, educational testing service; Seang-Hwane Joo, Educational Testing Service; Frederic Robin, Educational Testing Service; Kentaro Yamamoto, Educational Testing Service*

Validation of Extreme Response Style versus Rapid Guessing in Large-Scale Surveys *Ulf Kroehne, DIPF | Leibniz Institute for Research and Information in Education, Germany; Lale Khorramdel, National Board of Medical Examiners; Frank Goldhammer, DIPF | Leibniz Institute for Research and Information in Education, Centre for International Student Assessment (ZIB), Germany; Matthias von Davier, National Board of Medical Examiners*

Examining the Releation between Measurement Invariance and Response Styles in Cross-Country Surveys *Artur Pokropek, Educational Research Institute (IBE), Warsaw, Poland; Lale Khorramdel, National Board of Medical Examiners*

Session Organizers:

Lale Khorramdel, National Board of Medical Examiners
Artur Pokropek, Educational Research Institute (IBE), Warsaw, Poland
Janine Buchholz, Leibniz Institute for Research and Information in Education (DIPF)

Chair:

Lale Khorramdel, National Board of Medical Examiners

Discussant:

Leslie Rutkowski, Indiana University Bloomington

157. When Measurement Meets Causal Inference: Making Both Count

Coordinated Session *10:35 to 12:05 pm*

Hilton San Francisco Union Square: Continental 1

Exchanging views between different disciplines can provide new perspectives in each field. This symposium introduces five interdisciplinary studies that connect the measurement modeling framework and causal inference framework. Each presentation will illustrate how insights and techniques from one field (i.e., causal inference or measurement) can be valuable to better understand important issues in the other. The first paper shows how clear mathematical definitions of causal effects can resolve the long-standing debate on answer changing effects. The second paper formalizes rater effects with potential outcomes and demonstrates some advantages that can overcome the limitations of the conventional item response theory approach. The third paper discusses the importance of measurement error for causal effect estimation and why and how researchers' intuition on the impact of measurement error on bias can fail. The fourth paper then investigates whether and how measurement techniques such as confirmatory factor analysis can be used to improve causal effect estimation. Finally, the fifth paper argues the importance of causal reasoning in choosing a measurement model.

Participants:

Two Misconceptions About Answer Changing and Reviewing Effects in Multiple-Choice Exams *Yongnam Kim, University of Missouri, USA*

Framing Rater Effects as Causal Effects *Hyo Jeong Shin, educational testing service*

Does a Correlated Covariate Compensate Attentuation Bias of a Fallible Covariate *Marie-Ann Sengewald, Otto-Friedrich-Universität Bamberg, Germany; Steffi Pohl, Freie Universität Berlin*

The Role of Factor Analytic Techniques in Controlling for Confounding *Rui Lu, Teachers College, Columbia University, USA; Bryan Keller, Teachers College, Columbia University, USA*

The Importance of Causal Reasoning When Choosing a Measurement Model *Mijke Rhemtulla, University of California, Davis, USA*

Session Organizers:

Yongnam Kim, University of Missouri, USA
Hyo Jeong Shin, educational testing service

Discussant:

Daniel McCaffrey, Educational Testing Service

158. Differential Item Functioning #2

Paper Session *10:35 to 12:05 pm*

Hilton San Francisco Union Square: Continental 2

Participants:

Avoiding Omitted Variable Bias in DIF Assessment: A Regularized MIMIC Approach *Jyun-Hong Chen, Soochow University; Hsiu-Yi Chao, Academia Sinica*

The effect of OVB has long been ignored in DIF assessment. Consequently, the estimated DIF effects may not be unbiased, resulting in mistakenly DIF flagging. The lasso MIMIC method employed in the study successfully solves the problem in terms of well-controlled Type I error rates and high-power rates.

Dissecting Ability in DIF Analysis *Rabia Esma Sipahi Akbas, University of Kansas; John Poggio, University of Kansas*

Measuring true ability is paramount to DIF analyses. We propose measuring ability using examinee scores in different content areas (rather than total) and explore generalizability for impacted groups. Using NAEP 2015 assessments, we discover that defining true ability more exactingly yields more accurate race, gender, and ELL rooted DIF results.

Performances of Information Criteria as DIF Detection Methods with Variance Heterogeneity *Yong Luo, Educational Testing Service; Xinya Liang, University of Arkansas*

Ability variance heterogeneity has been shown to cause inflated Type I error with many differential item function (DIF) detection methods. The current research compares the statistical performances of several common information criteria (IC) indices as DIF detection methods with that of likelihood ratio test (LRT) with variance heterogeneity.

Robustness of Weighted Differential Item Functioning (DIF) Statistics *Ru Lu, Educational Testing Service; Hongwen Guo, Educational Testing Service; Neil Dorans, Educational Testing Service*

Many studies investigated the observed-score-based DIF statistics and found differences from their latent-ability-based counterparts. Recent theoretical studies show that using weighted sum scores as the matching variable can close the gap between the two. In this study, using simulated data, we evaluated the effectiveness of such modified DIF statistics.

The Relationship Between Reliability and DIF Detection Methodology *Jinmin Chung, University of Iowa; Ye Ma, University of Iowa; Terry Ackerman, University of Iowa*

This study examines how reliability of a test affects DIF detection methodology. Four different DIF detection procedures, SIBTEST, Mantel-Haenszel, Raju's DFIT, and Lord's Chi-square, were evaluated for six different levels of reliability, two different test lengths, uniform and non-uniform DIF, and four levels of sample size.

Chair:

Qi Diao, ETS

Discussant:

Sarah Quesen, Pearson

159. CAT Item Selection and Item Types

Paper Session *10:35 to 12:05 pm*

Hilton San Francisco Union Square: Continental 3

Participants:

Measurement Efficiency for Technology Enhanced and Multiple Choice Items in a CAT *Ozge Ersan, University of Minnesota; Yufeng Chang Berry, Minnesota Department of Education*

Multiple choice and four technology-enhanced item types (complex hot spot, graphic gap match, incline choice, fill in the blank) were analyzed for item information, and measurement efficiency. Results were examined across different grades and cognitive complexity in a K-12 online state accountability computerized adaptive mathematics test.

Penalization to Item Selection Inadequacy in a CAT *Chen Li, Kaplan Test Prep; Michael Chajewski, Kaplan Test Prep*

This study proposes the Sequence-Level Item Selection Efficiency Index (SISEI) as a modification of the Qualify of Item Pool Index (Gonulates, 2019). It penalizes selecting items consistently larger or smaller than the ability on consecutive positions, especially when the ability-difficulty discrepancy is large relative to the localized item selection range.

The Application of Multiple Imputation-Based Item Selection Rules in Computerized Adaptive Testing *YE FENG, Fordham University; Leah Feuerstahler, Fordham University*

Applications of computerized adaptive tests (CATs) typically ignore uncertainty associated with item calibration. This study proposes CAT item selection methods that incorporate item parameter uncertainty using a multiple imputation approach. Preliminary results suggest that less biased trait estimates with more accurate standard errors using the proposed methods versus tradition methods.

Using Machine Learning to Administer Salt Items in Computerized Adaptive Testing *Zhongmin Cui, ACT, Inc.; Chunyan Liu, National Board of Medical Examiners; Yong He, ACT*

Computerized Adaptive Testing with Salt (CATS) implements CAT with unrestricted item review and answer changes while being robust to cheating strategies. A successful implementation of CATS depends on effective administration of salt items to the right test takers. Machine learning was found to be helpful on improving the effectiveness.

Utilizing Item Response Time in Item Selection in Variable-Length CAT *Shuangshuang Xu, University of Maryland; Hong Jiao, University of Maryland*

To facilitate the computational efficiency of a variable-length CAT design, we adopt the per-time-unit maximum information method in item selection process. Also, a response-time-based joint-modeling Bayesian approach uses an informative empirical prior to estimate the ability parameter, and thus is instrumental to the termination of a variable-length CAT.

Chair:

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

Discussant:

Melinda Montgomery, College Board

160. Factor Analysis

Paper Session *10:35 to 12:05 pm*

Hilton San Francisco Union Square: Franciscan Ballroom A

Participants:

Adaptive Quadrature Estimation on Confirmatory Factor Analysis *Nermin Kibrislioglu Uysal, Hacettepe University; Kubra Atalay Kabasakal, NO; Burcu Atar, Hacettepe University*

This study purposes to examine the performance of adaptive quadrature estimation (AQ) method for confirmatory factor analysis. Specifically, we compared four link functions of AQ method under different CFA models, measurement intervals and differentiating degrees of nonnormality. The method is demonstrated via a simulation study.

Determining the Number of Factors with the Existence of Trivial Model-data Misfit *Yan Xia, University of Illinois at Urbana-Champaign*

This study investigated how parallel analysis methods, minimum average partial, Kaiser's rule, and sequential chi-square and RMSEA methods performed when population models could only be approximated by close-fitting models. Results revealed that Horn's parallel analysis could most accurately detect the number of factors given the existence of trivial misfit.

Evaluating the measurement of quantitative reasoning using items from domain-specific knowledge tests *Susanne Schmidt, Johannes Gutenberg University Mainz; Olga Zlatkin-Troitschanskaia, Johannes Gutenberg-University Mainz; Richard J. Shavelson, Stanford University*

We aimed to assess quantitative reasoning (QR) within the domain of business and economics (B&E) by asking whether it is possible to extract QR as the skill to handle quantitative data in existing tasks of B&E knowledge tests. Within a confirmatory factor model, QR items constitute a separable reliable factor.

Parallel Bifactor Analysis *Wes Bonifay, University of Missouri; Mark Hansen, University of California, Los Angeles*

Recent research has demonstrated that the bifactor model has a high propensity to fit well. This study improves upon traditional goodness-of-fit statistics by evaluating a hypothesized bifactor structure relative to many alternative structures. This approach facilitates detection of equivalent (or even better-fitting) structures, thereby providing stronger evidence for model selection.

Posterior Predictive Model Checks using Bayesian Saturated Models in Confirmatory Factor Analysis *Jihong Zhang, University of Iowa; Jonathan Templin, University of Iowa; Catherine Elizabeth Mintz, University of Iowa*

This study aims to propose a new method of posterior predictive model checking for Bayesian confirmatory factor analysis (CFA) models. Results show that the saturated model PPMC approach was an accurate method of determining local model misfit and could be used for model comparison.

Chair:

Yu Fang, Law School Admission Council

Discussant:

Ruben Castaneda, College Board

161. Multi-Stage Adaptive Testing

Paper Session *10:35 to 12:05 pm*

Hilton San Francisco Union Square: Franciscan Ballroom C

Participants:

An Investigation of the Impact of Item Pool Characteristics on Dynamic MST *Hyun Joo Jung, University of Massachusetts Amherst*

This study explores the impact of item pool characteristics, such as pool sizes and item difficulties, on dynamic multistage testing (dy-MST) which is a recently proposed adaptive testing method by Luo and Wang (2019). We examined how different item pool characteristics affect measurement precision and classification accuracy of dy-MST.

An Investigation on Pretest Item Recovery in Multistage Adaptive Testing *Rabia Karatoprak Ersen, University of Iowa; Won-Chan Lee, University of Iowa*

This study aims to evaluate impact of several variables on pretest item recovery in a 1-3 MST design when the pretest items are administered together with operational items. Fixed parameter calibration and separate calibration with linking are compared under each of 1PL, 2PL and 3PL IRT models separately.

Comparison of Two Adaptive Sampling Designs for Calibrating Multistage CAT Pretest Items *Shu Jing Yen, Center for Applied Linguistics; Kyoungwon Bishop, WIDA at WCER at UW-Madison*

An essential part of maintaining a multistage adaptive CAT program is to replenish test items through pretesting, however, there is a lack of research in designing pretest and in selecting sample for item calibration. This study introduced and compared two innovative adaptive sampling designs for calibrating multistage CAT pretest items.

Sequential or Simultaneous: Scale Linking for the Multistage Test *Tsung-Han Ho, ETS*

Simultaneous linking can be an efficient approach for scale linking in multistage test when numerous chain-linked tests are administered concurrently. The performance of simultaneous linking is evaluated by the comparison with mean/mean, mean/sigma, and Stocking-Lord procedure in terms of item parameter recovery and the stability of transformation across test conditions.

The Impact of Routing Strategy on Equated Number Correct Scores of an MST *Hacer Karamese, University of Iowa; Won-Chan Lee, University of Iowa*

The purpose of this simulation study is to assess the impact of routing strategy on equated number-correct scores under a three-stage MST design. Simulations are performed to compare routing methods, routing strategies, and equating methods.

Chair:

Eric William Shannon, Community College of Philadelphia

Discussant:

Samuel Haring, ACT

150. Electronic Board. Monday 10:35

Electronic Board Session *10:35 to 12:05 pm*

Hilton San Francisco Union Square: Salon A

Participants:

A Comparison of Item-Selection Methods in Certification/Recertification Testing *QIAO LIN, University of Illinois at Chicago; John Weir, NCCPA; Drew Dallas, NCCPA; Joshua Goodman, National Commission on Certification of Physician Assistants; Fen Fan, NCCPA*

This study investigates the performance of four methods of identifying and selecting remediation items that are based on examinees' responses to items and survey questions regarding confidence and relevance in a large-scale, longitudinal, certification/recertification exam. The effectiveness of the four methods are compared in terms of measurement precision and exposure control.

A Lognormal Response Time Model to Identify Examinees with Item Preknowledge *Murat Kasli, University of Miami; Cengiz Zopluoglu, University of Miami*

In this study, deterministic gated lognormal response time model was proposed to differentiate aberrant response behavior. The efficacy of the new model was demonstrated through simulation by manipulating the variables of sample size, number of items, percentages of compromised items and examinees with item preknowledge.

A Model for Multiple-Choice Exams based on Signal Detection Theory *Lawrence Thomas DeCarlo, Columbia University*

A signal detection model for multiple choice exams is developed. The model is a probabilistic mixture of nonlinear models and generalizes an earlier forced-choice model. The software Stan is used for Bayesian estimation with both MCMC and variational Bayes. The model is illustrated with simulated and real-world data.

Adaptive Statistics to Detect Aberrant Behavior in Testing *Igor Himelfarb, National Board of Chiropractic Examiners; Guoliang Fang, Penn State University; Andrew R Gow, NBCE*

Aberrant behavior during high-steak exams constitutes a threat to the validity of test scores. This paper presents an adoptive statistic development for detecting unlikely similar patterns of responses between pairs of examinees. Comparison of the results with findings provided by a commercial company showed better performance of the statistic.

Adverse Consequences of Early Semester Classroom Assessment for Developmental Community College Students *Charles Secolsky, Measurement and Assessment Consultant; Peter Arvaniites, Rickland Connubity College; Steven Holtzman, Educational Testing Service; Sathasivam Kriushnan, California State University Bakersfield; Eric Magaram, Rockland Community College; Matthew Macovich, Rockland Community College*

First classroom assessments administered to developmental community college mathematics students on the second/third class earned significantly lower final grades than students assessed the first time on the seventh/eighth class. Students who tested earlier likely experienced failure before "bonding" with instructor and classmates, leading to absenteeism, motivation decline, and inattentiveness.

An Approach to Assess Score Scale Drift under IRT Framework *Yanlin Jiang, ETS*

The study explores a flexible and feasible method to assess score scale stability when various drift conditions occur. It is particularly useful for test programs offering continuous testing. Simulated data are used for the study. The predicted score change based on the new approach are provided and evaluated.

An Attribute-level Item Exposure Control Method for DCM-CAT *Yu Bao, James Madison University; Laine Bradshaw, University of Georgia*

Empirical studies show diagnostic assessments are likely to have unequal attribute information. For an unbalanced item pool, this study proposes an attribute level item selection method with exposure control. We investigated classification accuracy and item exposure measure for the new method through a simulation study under various conditions.

Assessing QC of Large-Scale Assessment with Multiple Forms Using Harmonic Regression *Shuhong Li, ETS; Jiahe Qian, Educational Testing Service*

Harmonic regression was used to conduct quality control for an English-language assessment and to investigate the regional and year effects. For regional effect, a less conservative significance test with the Bonferroni correction was proposed given a shorter list of the retained variables through the backward selection procedure.

Bayesian analysis of Structural Equation Models with Nonignorable Missing Ordered Categorical Data *Jihang Chen, 8573165629; Zhushan Mandy Li, 2174935185*

A Bayesian approach for analyzing nonlinear structural equation models with nonignorable missing data is proposed to investigate the relationships among factors that impact student interest in science. This approach incorporates missing data mechanism into the model, thus avoids the biased results produced by conventional methods due to nonignorable missing data.

Comparing Item Response Prediction based on Machine Learning and Explanatory Item Response Theory *Konstantinos Pliakos, KU Leuven; Seang-Hwane Joo, Educational Testing Service; Jung Yeon Park, KU Leuven; Frederik Cornillie, KU Leuven; Celine Vens, KU Leuven; Wim Van den Noortgate, KU Leuven*

In educational measurement, missing item responses often occurs in online learning assessment. This study compares item response prediction accuracies based on Random Forests and IRT. We found that both approaches can be candidate methods to predict those missing item responses and possible solutions for constructing a valid assessment tool.

Does Measurement Matter? Connecting Issues of Measurement to Accuracy of Impact Estimates *Andrew Peter Jaciw, Empirical Education inc.; Thanh Nguyen, Empirical Education inc.; Li Lin, Empirical Education inc.*

Sensitivity analyses in impact evaluations typically assess how impact results vary depending on compositional and structural factors on the "right-hand side" of the equation. We leverage item-level data for an assessment to illustrate how the "left-hand side", mainly approaches to scaling assessments, may also affect the stability of impact estimates.

Exploring Factors and Effects of Social Media Fatigue *Shiyi Zhang, Beijing Normal University; Tao Xin, Beijing Normal University, PRC.*

The present study first found out a solution to the disaccord between assumptions and results in Bright's study. Then based on Stressor-Strain-Outcome frame of social media fatigue (SMF), the study added upward social comparison to stressor part and self-esteem to outcome part to further

explore factors and effects of SMF.

Functional clustering for detecting within-person dimensionality *George Engelhard, The University of Georgia; Victoria Tamar Tanaka, The University of Georgia; Kyle Turner, The University of Georgia*

Good model-data fit is necessary for measurement invariance. Although item fit is routinely assessed in psychometric studies, person fit is less often examined. We suggest the use of functional data analysis and functional clustering for estimating and interpreting person response functions.

Impact of Missing Documents on Latent Dirichlet Allocation Analysis *Minju Hong, University of Georgia; Allan Cohen, University of Georgia*

This study aims to investigate the effects of missingness on estimation of topic models using latent Dirichlet allocation. Preliminary results suggest that the number of topics extracted changed, but the top-10 words for each topic was unaffected.

Investigation of Social Desirability Bias in Self-Reported Measures Using Item Response Models *In-Hee Choi, Korean Educational Development Institute; Heekyung Kwon, Korean Educational Development Institutue*

This study investigates the effects of social desirability bias (SDB) in self-reported measures using item response models (IRMs). Two ways of differential item functioning (DIF) analysis, manifest and latent group DIF, were employed to explore possibility that SDB affected item responses. An empirical data from large-scale survey was analyzed.

Multilevel Analyses of Subgroups on the Major Field Test of Business *Guher Gorgun, University at Albany, SUNY; Kimberly Colvin, University at Albany, SUNY; Katrina Crotts Roohr, Educational Testing Service; Guangmin Ling, Educational Testing Service*

This study aims to investigate subgroup differences in a domain-specific standardized outcomes assessment used in higher education. Using multilevel modeling, we investigated the Major Field Test (MFT) for Business to determine whether, and to what extent, race/ethnicity or gender were related to the MFT test scores.

Predicting High School Advanced College Courses Success: State Assessments or PSAT *Jianlin Hou, The School District of Palm Beach County; Donghai Xie, The School District of Palm Beach; Paul Houchens, The School District of Palm Beach*

Since 2015, the state assessment and PSAT have some revolution, this study investigates how well do the state assessment and PSAT predict high school advanced college course success such as AP and AICE. It also evaluated the predication difference by race and gender.

The Impact of Different Missing Data Patterns on Longitudinal Diagnostic Classifications *Jiajun Xu, University of Georgia; Laine Bradshaw, University of Georgia*

Longitudinal diagnostic classification models can evaluate student knowledge components over time, identifying students' strengths and weaknesses throughout an extended lesson, curricular unit, school year, or other period of time. Ubiquitous missing data in longitudinal settings threatens measurement invariance. We investigate how different missing patterns affected the performance of longitudinal DCMs.

The robustness of Latent class analysis to violations of local independence assumption *Jungkyu Park, Kyungpook National University; Kwanghee Jung, Texas Tech University; Jaehoon Lee, Texas Tech University*

The purpose of this study is to examine the robustness of LCA parameter estimation and fit statistics in the presence of local dependence among indicators. A simulation study was conducted to identify what degree a local dependence could yield adverse impacts on accuracy in parameter estimation and model selection.

Using Latent Class Analysis for Early Phase Item Analysis *Sydne T McCluskey, CUNY Graduate Center; Magdalen Beiting-Parrish, CUNY Graduate Center; Jay Verkuilen, CUNY Graduate Center; Howard Everson, City University of New York; Claire Wladis, Borough of Manhattan Community College*

This paper introduces the use of latent class analysis (LCA) during early phase pilot testing - particularly when those analyses are conducted on innovative items which are designed to be multidimensional or are not well understood and traditional approaches to item analysis are therefore infeasible.

National Council on Measurement - ITEMS Module 4 *Andre Alexander Rupp, Educational Testing Service (ETS)*

Learn about your opportunity to publish in the ITEMS series. The ITEMS portal is your entry point into a world of learning in educational measurement and assessment. ITEMS modules are its centerpiece, which are short self-contained lessons with various supporting resources that facilitate self-guided learning, team-based discussions, as well as professional instruction.

163. **Invited Session Testing Irregularities - Real World Solutions**

Coordinated Session *12:25 to 1:55 pm*

Hilton San Francisco Union Square: Continental 4

Join us for a fun and interesting session focused on handling of testing emergencies. 15 minutes prior to the test session, our brave and courageous panelists will be emailed a testing scenario gone wrong. Minutes prior to the session, three teams will get together and map out an immediate response action plan and present to our panel of experts. Come here what went wrong, and how our panel recommends moving forward.

Presenters:

Andrew C. Dwyer, American Board of Pediatrics
Amin Saiar, PSI Services
Molly Faulkner-Bond, WestEd
Ben Domingue, Stanford University
Ricardo Mercado, Data Recognition Corporation
Jennifer Dunn, Questar Assessment

Session Organizer:

Andrew Wiley, ACS Ventures, LLC

Discussant:

Chad Buckendahl, ACS Ventures, LLC
Diane Henderson, ACT, Inc.

164. Enhancing Instruction and Testing Practice Using Natural Language Processing & Learning Analytics

Coordinated Session *12:25 to 1:55 pm*

Hilton San Francisco Union Square: Imperial Ballroom A

Technology affords analyses of student data that increase our understanding of student learning and inform enhancements in instruction and testing practice. Natural language processing (NLP) technology and learning analytics support large-scale data explorations. Through four presentations using the domains of writing and collaboration as illustrations, we discuss how NLP and learning analytics can support the exploration of student behavior toward learning, instruction and testing enhancements. NLP generates linguistic feature data from student data; learning analytics supports the analysis of data to understand and promote learning. The presentations will address these research questions: (1) How do NLP features from student writing on timed writing assessments compare to coursework writing?; (2) How can process and product data from writing tasks be used as formative feedback? ; (3) Can we identify meaningful patterns of use from process data that differentiate writers as they draft text in a writing app, and do these patterns relate to text change?; and, (4) How can we measure learning in collaborative learning environments, educational simulations, and intelligent tutoring systems? Using the domains of writing and collaboration, these presentations illustrate how NLP and learning analytics help us to study student behaviors to enhance future learning, instruction and testing practice.

Participants:

Are Standardized Writing Assessments Representative of Students' Writing? *Daniel McCaffrey, Educational Testing Service; Jill Burstein, Educational Testing Service; Beata Beigman Klebanov, Educational Testing Service; Guangmin Ling, Educational Testing Service; Steven Holtzman, Educational Testing Service*

Providing Formative Feedback on Writing Performance to Support Test Preparation *Mo Zhang, ETS; Chen Li, Educational Testing Service; Sandip Sinharay, ETS; Hongwen Guo, Educational Testing Service*

Uncovering Patterns of Use in the Writing Mentor® App: A Network Analysis Approach *Mengxiao Zhu, Educational Testing Service; Jiangang Hao, Educational Testing Service; Jill Burstein, Educational Testing Service*

Measurement Models for Collaborative Learning *Alina von Davier, ACTNext; Benjamin Deonovic, ACTNext by ACT; Michael Yudelson, ACTNext by ACT; Pravin Chopade, ACTNext by ACT; David Edwards, ACTNext by ACT*

Session Organizers:

Mo Zhang, ETS
Jill Burstein, Educational Testing Service

Chair:

Mo Zhang, ETS

Discussant:

Danielle McNamara, Arizona State University

165. College Admissions: Lessons Learned from Across the Globe

Coordinated Session *12:25 to 1:55 pm*

Hilton San Francisco Union Square: Imperial Ballroom B

In this session, an international group of experts on higher education admissions practices share their insight on opportunities and challenges related to the processes and criteria used in postsecondary admissions decision-making to promote access, equity, and fairness for candidates from diverse backgrounds. The session brings outside voices into the educational measurement community to engage in meaningful discussions about current and future-looking uses of assessments utilized to inform admissions practices, and discusses opportunities and challenges to developing culturally-responsive assessments that are sensitive to the ways of knowing and learning of diverse populations. The presenters discuss challenges in improving diversity, access, and equity in admissions processes used across the globe. The session uses a panel format to bring together voices from educational and professional communities and invites them to discuss various perspectives regarding access issues and challenges to diversifying the admitted student pool. The panel members will discuss their perceptions of what are the most critical measurement-based issues facing higher education admissions in their own country, why it is important to consider that perspective as part of fairness and access to admissions decision-making practices, and possible strategies to address the issue based on the lessons learned from their own country-level perspective.

Participants:

An Overview of Higher Education Admissions Processes *Rochelle Michel, ERBLEARN*

Access, Equity & Admissions Processes in South African Higher Education *Naziema Jappie, University of Cape Town*

Perspectives on Admissions Practices: The Case of Chilean Universities *Monica Silva, Pontificia Universidad Católica de Chile*

Character-Based Admissions Criteria: Validity and Diversity *Rob Meijer, University of Groningen*

Session Organizer:

Maria Elena Oliveri, Educational Testing Service

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

Coordinated Session *12:25 to 1:55 pm*

Hilton San Francisco Union Square: Yosemite A

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

Participants:

Willingness to Engage with a Low-Stakes Assessment: Evidence from a Natural Experiment in PIAAC *Francesca Borgonovi, Department of Social Science, Institute of Education, University College London, United Kingdom; Francois Keslair, 2Directorate for Education and Skills, Organisation for Economic Co-operation and Development, France; Marco Paccagnella, Directorate for Education and Skills, Organisation for Economic Co-operation and Development, France*

Exploring Group Differences in Large-Scale Assessments Using Latent Class Analysis on Process Data *Daniella Alves Rebouças, University of Notre Dame; Qiwei He, Educational Testing Service; Xiang Liu, Educational Testing Service*

Effects of Age in Dealing with Information from Search-Engine Environments: Results from an Analysis of PIAAC Log File Data *Carolin Hahnel, DIPF | Leibniz Institute for Research and Information in Education, Centre for International Student Assessment (ZIB), Germany; Frank Goldhammer, DIPF | Leibniz Institute for Research and Information in Education, Centre for International Student Assessment (ZIB), Germany; Ulf Kroehne, DIPF | Leibniz Institute for Research and Information in Education, Germany*

Clustering Behavioral Patterns Using Process Data in PIAAC Problem-Solving Items *Qiwei He, Educational Testing Service; Dandan Liao, American Institutes for Research; Hong Jiao, University of Maryland*

Session Organizer:

Qiwei He, Educational Testing Service

Chairs:

Qiwei He, Educational Testing Service

Frank Goldhammer, DIPF | Leibniz Institute for Research and Information in Education, Centre for International Student Assessment (ZIB), Germany

Discussant:

Matthias von Davier, National Board of Medical Examiners

167. Alignment Frameworks for Complex Assessments: Score Interpretations Matter

Coordinated Session *12:25 to 1:55 pm*

Hilton San Francisco Union Square: Yosemite B

The design and validation of an assessment system, intended for both formative and summative purposes, requires careful development processes especially when such assessments are intended to support interpretations regarding how student learning grows more sophisticated over time. Under a principled approach to test design, the intended test score interpretation is defined, the evidence needed to draw a conclusion about where a student is in their learning based on that interpretation is defined, and items are developed according to those evidence pieces. The assessment of complex constructs such as student learning of NGSS and college and career standards may mean that traditional alignment and validity evidence is no longer optimal evidence that a test is aligned to state standards and its purpose. This session will focus on emerging frameworks for alignment and validity evidence explicitly designed to ensure that the assessment development process and evidence collection is cohesively centered in score interpretation. Experts in achievement level descriptors, alignment, principled assessment design, and standard setting will share emerging methodologies that fuse separate and previously distinct activities of test development, so these activities are embedded together into a cohesive whole in which score interpretations centered in student learning are the central focus.

Participants:

Evaluating Alignment Between Complex Expectations and Assessments Meant to Measure Them *Ellen Forte, edCount, LLC*

Examining Alignment of Test Score Interpretations on a Computer Adaptive Assessment *M. Christina Schneider, NWEA; Mary Veazey, NWEA*

Examining Alignment of Test Score Interpretations Using Multiple Alignment Frameworks and Multiple Measures *Karla Egan, EdMetric*

Embedded Standard Setting: Standard Setting as a Resolution of the Alignment Hypothesis *Daniel Lewis, Creative Measurement Solutions; Robert Cook, ACT, Inc*

Session Organizer:

M. Christina Schneider, NWEA

Discussant:

Paul Nichols, NWEA

168. d Assessment of Mathematical and Scientific Reasoning: An alternative to Machine-scoring Open-Ended Items

Coordinated Session *12:25 to 1:55 pm*

Hilton San Francisco Union Square: Yosemite C

This symposium will be an opportunity to hear about 3 projects investigating the assessment of mathematical and scientific thinking and the nature and quality of the feedback these projects provide. Although the topic of machine-scoring of open-ended item responses is a popular one today, the focus in this symposium will be on an alternative to machine-scoring. In common with (supervised) machine scoring, this approach starts by gathering open-ended responses to well-designed and topic-relevant prompts, and proceeds by developing scoring guides for those open-ended responses, and scoring the accumulated responses. However, instead of using these materials as training materials, in these 3 projects, we have been using those materials to develop technology-enhanced items that do not require students to write down their responses. One of the aims is to come up with assessments that are not limited by students writing skills. The 3 projects report varying levels of success, and each discusses the reasons why, and proposes steps that are aimed at improving the technique. The symposium provides an opportunity to learn about both the success and challenges of this approach and to contrast the strengths and weaknesses of each project. The projects are at the middle-school, high-school and college levels.

Participants:

Investigating an Alternative to Machine-scoring of Open-Ended Items *Mark Wilson, University of California, Berkeley*

Developing selected response items for a data-based decision-making instrument *Amy Arneson, Education Northwest*

An Exploration of Selected-Response Items Compared to Constructed-Response Item Types in Science Education *Linda Morell, University of California, Berkeley; Weeraphat Suksiri, University of California-Berkeley; Sara Dozier, Stanford University; Jonathan Osborne, Stanford University; Mark Wilson, University of California, Berkeley*

Comparing Selected Response and Constructed Response Items in Mathematical Problem-Solving *Yukie Toyama, University of California; Jerred Jolin, University of California, Berkeley; James Mason, University of California, Berkeley; Mark Wilson, University of California, Berkeley*

Session Organizer:

Mark Wilson, University of California, Berkeley

Chair:

Karen Draney, University of California Berkeley

Discussants:

Richard Patz, University of California-Berkeley

Jonathan Osborne, Stanford University

169. Identifying and Addressing Inauthentic Response Strategies in Automated Scoring

Coordinated Session *12:25 to 1:55 pm*

Hilton San Francisco Union Square: Continental 1

Long before multiple-choice became the workhorse item format, large-scale testing programs made use of essay, spoken word, and short answer formats with human raters for scoring. As technology has advanced, we have gained the ability to score these responses using computer algorithms, reducing the cost and increasing the availability of these formats. However, as automated scoring has moved into operational use for high-stakes applications (e.g., employment, immigration, school promotion), some test takers have explored strategies to increase their scores when automated scoring is used. Behaviors range from strategies such as "never use a simple word when a complex one could increase your score" to memorizing essay templates or stock phrases. This session will include 5 papers looking at strategies identified in testing programs, detection of anomalous behavior, and operational considerations for constructing and using automated scoring.

Presenters:

Susan Lottridge, AIR
Aoife Cahill, ETS
Xinhao Wang, ETS
William Bonk, Pearson

Participants:

Test security, validity, and automated scoring *Kirk Alan Becker, Pearson*

Comparing Deep Learning versus Bag-of-Words Robustness to Gaming Strategies *Susan Lottridge, AIR; Amir Jafari, AIR; Christopher M Ormerod, American Institutes for Research*

Detecting incoherence in automatically generated essays *Aoife Cahill, ETS; Michael Flor, ETS; Martin Chodorow, CUNY*

Automatic Detection of Gaming Spoken Responses Based on Very Deep Convolutional Neural Networks *Xinhao Wang, ETS; Keelan Evanini, ETS; Su-Youn Yoon, ETS; Yao Qian, ETS; Klaus kzechner@ets.org, ETS*

Exploring test-taker strategies for gaming automatically scored foreign language test speaking items *William Bonk, Pearson; Mallory Klungtvedt, Pearson; Saerhim Oh, Pearson; Jooyoung Lee, Pearson*

Session Organizer:

Kirk Alan Becker, Pearson

Discussant:

Alistair Van Moere, MetaMetrics

170. Evaluating Psychometric Models

Paper Session *12:25 to 1:55 pm*

Hilton San Francisco Union Square: Continental 2

Participants:

Computation of S-X^2 Item-Fit Statistic *Hyung Jin Kim, The University of Iowa; Amy Hendrickson, College Board*

In computing S-X^2, there exist two perspectives regarding obtaining expected-values, multiple procedures for collapsing tables of observed- and expected-values, and various choices for minimum cell values needed for collapsing. By evaluating and comparing those factors, this study aims to provide practical implications about obtaining appropriate S-X^2 for its proper uses.

Detecting Aberrant Behavior in CAT: The Joint Model for Accuracy and Speed *Xiaowen Liu, University of Connecticut; H.Jane Rogers, University of Connecticut*

A person-fit statistic under the joint model for accuracy and speed is used to detect aberrant test-taker responses and response time patterns in CAT. A simulation study is designed to test the performance of the person-fit statistic. Results shows this person-fit statistic provides more information in identifying aberrances.

Out-of-sample Predictive Performance Metrics for Item Response Models *Ben Stenhaug, Stanford University*

We derive two expected log predictive likelihood metrics for item response models based on different definitions of out-of-sample. Our goal is to enable and encourage psychometricians to conceptualize model comparison in a way that is consistent with best practices in statistics and computer science.

Review and Guidelines for Bayesian Approximate Measurement Invariance: What to Report? *Abeer Alamri, The National Center for Assessment; Eunsook Kim, University of South Florida*

This systematic review aims to investigate how Bayesian approximate measurement invariance (BAMI) was implemented through the exploration of peer-reviewed studies conducted between 2012 and 2017. Review results stress the need to delineate guidelines to know how to utilize BAMI estimation method, convergence, evaluation, and what to report in methodology/results sections.

Using New Reference Bifactor Models to Overcome Shortcomings of Traditional Bifactor Models *Wei S Schneider, University of Iowa; Walter P. Vispoel, University of Iowa; MURAT KILINC, University of Iowa*

We compared traditional bifactor modeling to a new approach in which a reference factor replaces the general factor in traditional modeling. Results from multidimensional inventories of self-concept and socially desirable responding revealed that the new approach reduced shortcomings of traditional modeling while extracting new diagnostic information about response patterns.

Chair:

Jessalyn Smith, DRC

Discussant:

Goran Lazendic, ACER

171. Machine Learning Approaches

Paper Session *12:25 to 1:55 pm*

Hilton San Francisco Union Square: Continental 3

Participants:

A Deep Learning Approach for Student Performance Prediction *Shumin Jing, University of Iowa; Sheng Li, University of Georgia; Won-Chan Lee, University of Iowa*

This study presents a deep learning model called deep response prediction (DRP) for student performance prediction. By converting students, items, and skills into high-dimensional sparse vectors, a deep neural network is trained to predict responses from students to items. Evaluations on real datasets demonstrate the effectiveness of the proposed model.

Comparison of Natural Language Processing Methods of Matching Items to Learning Objectives *tanya longabach, Kaplan Professional*

The purpose of this study was to find an accurate way to match items to learning objectives (LOs) automatically for an assessment leading to a certificate in financial education. The study compares topic modeling (TM) and latent semantic analysis (LSA). We found that LSA was performing better than TM.

Personalizing High-stakes Assessments with Recommender Systems *Saed Qunbar, AdvancED | Measured Progress; Robert Furter, American Board of Pediatrics*

A recommender system was tested with pilot data then implemented with the launch of a longitudinal assessment program to balance individualized content and the breadth of content represented by the certificate. This talk outlines the studies completed on the pilot data and the planned study conceived to assess model outcomes.

Prediction of Reading Fluency Scores with Silence Tendencies Using Machine Learning Algorithms *Chalie Patarapichayatham, Southern Methodist University; Akihito Kamata, Southern Methodist University; Yusuf Kara, Southern Methodist University*

This study investigates evidence for meaningful pauses that are believed to be made by fluent readers in the context of oral reading fluency (ORF) assessments. We derive various variables from between-word silence relative to total reading time and fit machine learning algorithms to classify readers into groups of ORF levels.

Using Machine Learning Approaches to Optimize Screening for Mathematics Difficulties *Damien Cormier, University of Alberta; Okan Bulut, University of Alberta; Eric Stickney, Renaissance Learning*

The use of multiple screening measures may improve accuracy when identifying students at-risk for low math performance. We compared predictive models generated from machine learning algorithms to traditional screening approaches. Our results suggest that machine learning algorithms outperform traditional screening approaches when predicting math performance.

Chair:

Bradley McMillen, Wake County Public School System

Discussant:

Ye Tong, Pearson

172. Design of CAT Driven Assessments

Paper Session *12:25 to 1:55 pm*

Hilton San Francisco Union Square: Franciscan Ballroom A

Participants:

Effects of Splitting Within-Testlet Items for Testlet-Based Computerized Adaptive Testing *Unhee Ju, Riverside Insights; Rong Jin, Riverside Insights; JP Kim, Riverside Insights.com*

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

Fair Testing Time for All: Constructing Multistage Adaptive Tests Using Response Times *Yooyoung Park, University of Massachusetts Amherst*

This study proposes a new approach to construct MSTs to make total testing time equitable across examinees using the lognormal model for response times (van der Linden, 2006). How using response times in constructing MSTs interplays with MST design features and its resulting impact on test completion time are investigated.

Modified Constrained Adaptive Testing with Shadow Tests to Improve Sub-score Accuracy *Ann Hu, NWEA; Yuehmei Chien, NWEA*

The shadow test approach to computer-adaptive testing selects items to not only allow for optimal adaptation but also realize all constraints simultaneously. This study examines the performance of content balance by standard error of measurement in a variable-length adaptive testing using shadow test technique, specifically focusing on sub-score accuracy.

On the Design of Interim Cognitive Diagnostic CAT in Learning Context *Chun Wang, University of Washington*

The interim CD-CAT differs from the current available CD-CAT designs primarily because students' mastery profile (i.e., skills mastery) changes due to learning, and new attributes are added periodically. In this talk, I will discuss the several designs of interim CD-CAT that are suitable in the learning context.

Optimal Online Calibration Designs for Item Replenishment in Adaptive Testing *Ping Chen, Beijing Normal University; Yinhong He, Nanjing University of Information Science and Technology*

A new online calibration design (D-c) is proposed by incorporating the idea of D-optimal design into the reformed D-optimal design (van der Linden & Ren, 2015). To better handle the dependence of design criteria on the unknown item parameters, Bayesian D-c is put forward by adding prior to new items.

Chair:

Mark L Davison, University of Minnesota

Discussant:

Kristin Morrison, Curriculum Associates

173. Multidimensionality

Paper Session *12:25 to 1:55 pm*

Hilton San Francisco Union Square: Franciscan Ballroom C

Participants:

A relation between multi-group multidimensionality and uniform DIF *Saemi Park, The Ohio State University; Paul De Boeck, Ohio State University*

'Multidimensional model for DIF' explicates how additional latent traits play a role in creating uniform DIF by considering three factors: group mean difference in latent dimensions and interdimensional correlation. An interrelation among them defines a function, DIF potential. We investigate in what way DIF potential affects properties of uniform DIF.

Evaluating Noncompensatory MIRT Models for Passage-Based Tests *Nana Kim, University of Wisconsin-Madison; Dan Bolt, UW-Madison School of Education*

This paper introduces noncompensatory MIRT models for passage-based tests. The models emphasize the difficulty of the response subprocesses by attaching a separate difficulty parameter to each component. Their practical value against the bifactor MIRT model is demonstrated through a real data application to a reading comprehension test.

Practical Designs of Testlet-based CAT within a Multidimensional Data Structure *JING-RU XU, Pearson; Joe Betts, NCSBN*

This research explored different designs of testlet-based CAT under a practical testing context. Designs under different conditions were simulated. Evaluation criteria were computed to compare the efficiency and accuracy between different designs. The results enlighten the value of practical implications of the new design in a real multidimensional adaptive testing context.

Self-Report Rating Scale Format, Item Location, and Susceptibility to Response Style Effects *Dan Bolt, UW-Madison School of Education; Yang Caroline Wang, Education Analytics, Inc.; Robert Meyer, Education Analytics, Inc.*

We use a multidimensional IRT model for response style to examine how rating scale format may relate to examinee tendencies to adopt response styles. Rating scales that vary anchors across items appear more susceptible to response style effects, although item location also emerges as a factor.

The Design of Q Matrix for Multidimensional Diagnostic Models *Yiling Cheng, Kaohsiung Medical University, Taiwan; Mark Reckase, Michigan State University; Barbara Schneider, Michigan State University*

The primary purpose of the study was to examine the effect of Q matrices when a hierarchical structure is presented. Five designs of Q matrices were tested with TIMSS 2011 data. A simulation study was also conducted to examine the influences of different Q matrices on relative model fits.

Chair:

Aileen Reid, University of North Carolina at Greensboro

Discussant:

Scott Monroe, University of Massachusetts Amherst

162. Electronic Board. Monday 12:25

Electronic Board Session *12:25 to 1:55 pm*

Hilton San Francisco Union Square: Salon A

Participants:

A Literature Review of Interim Assessment Use with Implications for Interpretive Arguments *Nathan Dadey, Center for Assessment; Calvary Diggs, University of Minnesota*

To date there has not been a systematic review of assessments explicitly identified as 'interim'. The purpose of this study is to identify and summarize available literature on interim assessments and their uses, and then draw on the resulting synthesis to develop interpretive arguments based on specific, empirically supported uses.

A Vector Approach to Identifying Partially Overlapping Groups in a Multidimensional Space *Jonathan Weeks, ETS*

This study implements a vector approach for clustering individuals. The method relies on a comparison of reference composite vectors across multiple subgroups. Empirical data from a reading assessment are used to identify students with different levels of skill integration. The results are consistent with the strand model of reading development.

An Adaptive Method for the Online Calibration of Q-matrix for New Items in CD-CAT *Teng Wang, School of Computer Information Engineering, Jiangxi Normal University; Wenyi Wang, School of Computer Information Engineering, Jiangxi Normal University; Lihong Song, Jiangxi Normal University; Peng Gao, Jiangxi Normal University; Jian Xiong, Jiangxi Normal University*

The purpose of this study is to propose an adaptive online calibration method based on the Shannon entropy to select the most suitable new items for the examinees

Assessing Intervention Effects Using Two-Tier Item Factor Models *Yon Soo Suh, UCLA; Li Cai, University of California—Los Angeles*

This study explores the applicability of two-tier item factor models for testing intervention effects while accommodating various measurement issues often embedded within randomized experiments. The performance of the proposed approach is evaluated using an empirical data analysis as well as a simulation study. Practical implications are discussed.

Assessment reports aimed at designing academic interventions: a study conducted in India. *Prashanth Vasudevan, Gray Matters India; Perman Gochyyev, BEAR Centre*

In India, assessments are used for criteria validation and report test scores only. A well-designed assessment—analyzed using Item Response Theory—can provide insightful feedback for improving academic interventions. Here, we describe a large, first-of-its-kind study in India, in which assessment data was made actionable using reporting framework.

Characterizing Uncertainty in School Accountability Indicators Computed from Student Test Scores *Mark Hansen, University of California, Los Angeles*

School-level accountability indicators are often computed from student test scores but ignore the uncertainty in those scores. We propose using plausible values to characterize the effects of measurement error on accountability indicators. The approach is illustrated using a statewide sample of English Learners in grades 9-12.

Designing a Programmatic Approach for the Assessment of Competencies for Collaborative Practice *Mary Roduta Roberts, University of Alberta; Sharla King, University of Alberta; Iris Cheng In Chao, University of Alberta*

This collaborative study aimed to design an assessment plan for an institutional interprofessional education curriculum undergoing redesign. We will share a summary of the IP assessment resources developed, describe the continuum of programmatic assessment models, next steps in our collaboration with stakeholders, and reflections on our approach.

Exploring Socialization Practices on Educational Experiences and Academic Performance Using NIES Data *Stephanie Peters, Stephanie Peters; Emmanuel Sikali, Assessment Division of the National Center for Education Statistics, US department of Education*

The current research uses data from the National Indian Education Study (NIES). This study aims to assess the predictive power of culturally relevant socialization practices on the educational experiences and academic performance of a nationally representative sample of American Indian and Alaska Native (AI/AN) students. Regression analysis results are discussed.

Exploring the Estimation of Operational Item Statistics using Deep Neural Networks *Shichao Wang, ACT*

This study aims to investigate the application of deep neural networks to improve the prediction of operational item statistics by incorporating content and pretest item information. Observed data from test forms comprised of four subject areas were analyzed. The proposed method provided a reasonable prediction of operational item difficulties.

Fit Indices of Dynamic Bayesian Networks: A Comparison Based on PPMC *Yuxi Qiu, University of Florida*

Based on a Monte Carlo simulation, this study examined the performance of selected fit indices to detect misspecifications of dynamic Bayesian networks, an alternative psychometric modeling framework to portrait student learning in complex assessment settings, in the context of varied temporal dependence structure, sample size, and test length.

Impact of Item Misfit on Group Score Reporting in Large-Scale Assessments *Usama Ali, ETS; Seang-Hwane Joo, Educational Testing Service; Frederic Robin, Educational Testing Service*

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

Is student engagement in a low-stakes assessment equivalent across schools and countries? *EUN HYE HAM, Kongju National University; Hyo Jeong Shin, educational testing service; Yun Kyung Kim, Seoul National University*

This study investigates whether student engagement in a low-stakes international assessment was equivalent across schools as well as countries. Using the processing data from PISA 2015, the distributions of response time effort were compared in different conditions and also regressed by relevant school-level and country-level characteristics.

Is Three Still the "Optimal" Number of Response Options: A Meta-Analysis *Michael Dosedel, University of Minnesota*

Despite interest in innovative item types, common types, specifically single best response continue to dominate. This study applies modern techniques

in meta-analysis to the most researched item writing facet: how many response options are optimal for multiple choice questions.

Joint Modeling of Process and Response Data: A Mixture Markov Process Approach *Clifford Erhardt Hauenstein, Georgia Institute of Technology; Matthew Johnson, Educational Testing Service; Jie Gao, Educational Testing Service*

Using data from the 2015 NAEP Science Assessment, action sequences and responses are jointly modeled in a manner that characterizes between and within cluster variation in the response process. Specifically, a mixture distribution Markov model is applied to classes of actions; and cluster-specific initial, transition, and stationary probabilities are derived.

Measurement of Social and Emotional Competencies in Middle School *Feifei Ye, RAND Corporation; Catherine Augustine, RAND Corporation*

Using teacher-reported scale to assess middle school students' social-emotional competencies (SEC), this study examines whether teachers' rating is consistent while students receive multiple ratings from different teachers, whether teacher rating depends on course subject and teachers' familiarity with student, and how student SEC changes from primary school to middle school.

Racial Bias in Pre-Employment Tests *David Eugene Johnson, TransformEd Consulting*

Pre-Employment tests can violate anti-discrimination laws if used to discriminate by race. Social justice demands deconstructing phenomena to disclose ways society reproduces inequity. The Civil Rights Act permits employment tests if they do not discriminate. There appears to be no studies considering fairness of pre-employment tests. We must deconstruct them.

The Effects of Item Characteristics on lz and Ht Person-Fit Statistics *Yoon Jeong Jeong Kang, American Institutes for Research; Ming Li, Georgetown University*

This study investigates the effects of item characteristics on the lz and Ht person-fit statistics in detecting cheating examinees. The simulation results show that the empirical cutoff values and detection power of the lz and Ht statistics vary on item characteristics, particularly level of item discrimination in a test.

The Effects of Item Weighting on Scores in Mixed-Format Tests *Nathan Minchen, Pearson; Malena McBride, Pearson; Aimee Boyd, Pearson*

Weights can be used to adjust the contribution of individual items for scoring. Compelling reasons exist to support the use of weighting, but care must be taken to ensure that unintended consequences do not reduce the validity of the scores. This research explores conditions that impact the effect of weights.

Using Linear Mixed-effects Model to Detect Fraudulent Erasures at an Aggregate Level *Luyao PENG, ACT, inc; Sandip Sinharay, ETS*

Please see attached .pdf file

Using New Items to Detect Preknowledge Behaviors in Computerized Multistage Testing *YIBO WANG, University of Iowa; Deborah Harris, University of Iowa; Stephen Dunbar, University of Iowa; Lida Chen, University of Iowa*

Examinees may have preknowledge on reused items. We can detect them by comparing their performance on new and old items. This study investigates using the multistage testing under different patterns of new and over-exposed items and determines the effectiveness of methods in identifying examinees with pre-knowledge across the ability range.

National Council on Measurement - ITEMS Module 2 *Andre Alexander Rupp, Educational Testing Service (ETS)*

Learn about your opportunity to publish in the ITEMS series. The ITEMS portal is your entry point into a world of learning in educational measurement and assessment. ITEMS modules are its centerpiece, which are short self-contained lessons with various supporting resources that facilitate self-guided learning, team-based discussions, as well as professional instruction.

175. **Invited Session Computational Psychometrics as a Validity Framework for Process Data**

Coordinated Session *2:15 to 3:45 pm*

Hilton San Francisco Union Square: Plaza A

In 2015, von Davier coined the term "computational psychometrics" (CP) to describe the fusion of psychometric theories and data-driven algorithms for improving the inferences made from technology-supported learning and assessment systems (LAS). Meanwhile, "computational" [insert discipline] has become a common occurrence. In CP the process data collected from virtual environments should be intentional: we should design & provide ample opportunities for people to display the skills we want to measure. CP uses the expert-developed theory as a map for the measurement efforts using process data. CP is also interested in the knowledge discovery from the (little, big) process data. In this symposium, several examples of applications of computational models for the process data from learning systems and from the assessment of the 21st Century skills are presented. Psychometric theories and data-driven algorithms are fused to make accurate and valid inferences in complex, virtual learning and assessment environments.

Presenters:

Yuchi Huang, ACT
Lu Ou, ACT
John Whitmer, ACTNext

Session Organizer:

Alina von Davier, ACTNext

Discussant:

Bruno D. Zumbo, UBC

176. Separate but (Un)Equal? Measurement Expertise vs. Policy Knowledge in Testing for Accountability

Coordinated Session *2:15 to 3:45 pm*

Hilton San Francisco Union Square: Imperial Ballroom A

This proposal is for a moderated panel discussion for the 2020 NCME conference program. The session would fall under the "coordinated sessions" category in the NCME Call for Proposals. The panel's focus would be on the following theme: As Elementary and Secondary Education Act (ESEA) is slated for re-authorization again, this panel considers why assessment experts must also be policy experts – combining technical advice for K-12 test users and stakeholders with advice about the validity of the consequences of test use in accountability contexts.

Participants:

Does the Measurement Community 'Own' the Problems on Interpretation and Use of Test Scores? *Ellen Forte, edCount, LLC*

Ensuring that Test Use Leads to the Intended, Positive Consequences, while Minimizing Unintended, Negative Consequences *Suzanne Lane, University of Pittsburgh*

Towards Inclusive Assessments for all Students, and Inclusive Accountability Systems *Martha Lynn Thurlow, NCEO/University of Minnesota*

Mapping the Relationship Between Policy and Measurement Considerations in Supporting Accountability Testing *Neal Kingston, University of Kansas*

Separate and Unequal? Lessons from Selective High Schools Uses of Standardized Test Scores *Howard Everson, City University of New York*

Why Assessment Professionals Should Cultivate Policy Analysis Skills *Aaron Pallas, Columbia University, Teachers College*

Conflicting Theories of Action? Reconciling Mindsets of Measurement vs. K-12 Stakeholders to Improve Validity and Consequences of Testing *Madhabi Chatterji, Columbia University, Teachers College*

Session Organizer:

Madhabi Chatterji, Columbia University, Teachers College

177. Recent Advances in Research on Response Times

Coordinated Session *2:15 to 3:45 pm*

Hilton San Francisco Union Square: Yosemite A

Given the increasing popularity of computerized testing, the question of how to utilize response times has become urgent. This session provides a glimpse of the recent research using response times and suggests several new approaches involving response times. The approaches are demonstrated using data sets from high-stakes tests. The first three presentations focus on the use of response times to detect test fraud, detect low motivation, and improve test-form assembly. The last two presentations suggest new models involving response times. The session includes discussion from an expert in analysis of response times.

Participants:

Detecting Item Preknowledge Using Response Accuracy and Response Times *Sandip Sinharay, ETS; Matthew Johnson, Educational Testing Service*

A New Multi-Method Approach of Using Response Time to Detect Low Motivation *Ying Cheng, University of Notre Dame*

A Response Time Process Model for Not-reached and Omitted Items in Standardized Testing *Jing Lu, Northeast Normal University; Chun Wang, University of Washington*

Semi-parametric Factor Analysis for Response Times *Yang Liu, University of Maryland; Marian M Strazzeri, University of Maryland*

Session Organizer:

Sandip Sinharay, ETS

178. Integrating Timing Considerations to Improve Testing Practices

Coordinated Session *2:15 to 3:45 pm*

Hilton San Francisco Union Square: Yosemite B

This session highlights chapters from an upcoming book in the NCME Book Series, Integrating Timing Considerations to Improve Testing Practices. This exciting volume synthesizes research on the most significant topics in the field of timing in order to provide practitioners with a valuable source of practical advice to consider when making test design and scoring decisions. In addition to the five topics discussed in more detail below, additional topics covered in the book include: Timing considerations in test development and administration, extended testing accommodations for students with disabilities, comparison of experimental vs. observational approaches to evaluating the impact of time limits, timing considerations in simulations, games, and other performance assessments, impact of test timing on mode and device comparability, using response time for measuring cognitive ability, and use of response time and response accuracy for detecting item preknowledge.

Participants:

The Evolving Conceptualization and Evaluation of Test Speededness: A Historical Perspective *Daniel Jurich, National Board of Medical Examiners; Melissa Margolis, National Board of Medical Examiners*

The Impact of Time Limits and Timing Information on Validity *Michael Kane, Educational Testing Service*

Relationship between Testing Time and Testing Outcomes *Brent Bridgeman, ETS*

Response Times in Cognitive Tests: Interpretation and Importance *Paul De Boeck, Ohio State University; Frank Rijmen, American Institutes for Research*

A Cessation of Measurement: Identifying Test Taker Disengagement Using Response Time *Steven L. Wise, NWEA; Megan R. Kuhfeld, NWEA*

Session Organizer:

Arden Ohls, National Board of Medical Examiners

Chair:

Richard Feinberg, National Board of Medical Examiners

Discussant:

Richard Feinberg, National Board of Medical Examiners

179. Learning Metrics: Using Log Data to Evaluate EdTech Products

Coordinated Session *2:15 to 3:45 pm*

Hilton San Francisco Union Square: Yosemite C

In the past decade, technology-based assessment in education has evolved considerably. Major research syntheses of how people learn (NRC, 2000, 2018) recognize that digital technologies have the potential to support learners in meeting a wide range of goals in different contents. However, students who use these products generate an immense amount of data that can be difficult to interpret and use. Sophisticated measurement models are being developed to precisely describe what students know and can do in a digital environment. However, a more pressing need is to use log data to evaluate whether students are learning. This session consists of three perspectives on how EdTech companies use metrics related to learning to evaluate the effectiveness of their products. The first presentation describes how Duolingo uses log data to evaluate their language teaching approach and identify user behaviors that are associated with better learning outcomes. The second presentation discusses how Khan Academy combines log data with standardized test scores through district partnerships to evaluate its product. The third presentation describes how Digital Promise partners with teachers and school leaders to develop measures of learning from event data. These presentations will be followed by comments from an expert in learning technologies.

Presenters:

XIangying Jiang, Duolingo

Andrew E. Krumm, Digital Promise

Participants:

Learning Effectiveness and User Behavior: The Case of Duolingo *XIangying Jiang, Duolingo; Joseph Rollinson, Duolingo*

Measuring Classroom Learning in the Context of Research-Practice Partnerships *Rajendra Chattergoon, Khan Academy; Kodi Weatherholtz, Khan Academy; Kelli Millwood Hill, Khan Academy*

Measuring Learning Behaviors Using Data from a Common Learning Management System *Andrew E. Krumm, Digital Promise*

Session Organizer:

Rajendra Chattergoon, Khan Academy

Discussant:

David Porcaro, Chan Zuckerberg Initiative

180. Innovations in Detection of Test Collusion

Coordinated Session *2:15 to 3:45 pm*

Hilton San Francisco Union Square: Plaza B

This coordinated session aims to help close the gap between research and practice in detection of compromised groups, which remains as one of the most significant challenges in test security. This session brings together several leading scholars in the area of test security to introduce and discuss four novel approaches to detecting collusion. The papers align strongly with the conference themes to (1) leverage technology to (2) help improve assessment practices and (3) improve the fairness of educational assessments. Collectively, the set of papers will introduce new approaches which simultaneously detect compromised items and the groups of individuals engaged in collusion, leverage the latest technology and computational methods to improve detection through machine learning and the incorporation of eye tracking process data, and develop and investigate the first method to detect examinees with preknowledge during the exam and in real-time.

Participants:

Assessing pre-knowledge cheating via innovative measures: a multiple-group analysis of jointly modeling the item responses, response times, and visual fixation counts *Kaiwen Man, University of Maryland; Jeffrey Harring, University of Maryland; Cengiz Zopluoglu, University of Miami*

An Iterative Unsupervised-learning-based Approach for Detecting Item Preknowledge *yiqin pan, University of Wisconsin-Madison; James Wollack, University of Wisconsin*

A New Approach to Detection of Collusion: Iterative Cluster Building *James Wollack, University of Wisconsin; Sakine Gocer Sahin, WIDA at WCER at UW-Madison*

Real-Time Anomalous Response Detection for Computer-Based Linear Tests *Merve Sarac, University of Wisconsin-Madison; James Wollack, University of Wisconsin*

Session Organizer:

James Wollack, University of Wisconsin

Discussant:

Carol Eckerly, Educational Testing Service

181. Validity

Paper Session *2:15 to 3:45 pm*

Hilton San Francisco Union Square: Union Square 19/20

Participants:

Constructing a Validity Argument for Large-Scale Instructionally Embedded Assessments *Amy K Clark, University of Kansas; Meagan Karvonen, University of Kansas*

This paper describes the validation approach for a large-scale instructionally embedded assessment system. The validation process included articulation of claims in a theory of action, specification of underlying propositions, and evidence collected for each proposition. Evidence is combined into an argument evaluating the extent to which results support intended uses.

Designing Score Reports to Maximize Validity and Instructional Utility *Karen Barton, Edmentum; Audra Kosh, Edmentum*

Score reports are a critical point of validity, impacting test interpretation and use consequences. The study evaluates the impact of score reports to increase valid interpretations by comparing test administration practices and student effort, both of which can directly impact score comparability, before and after the score report redesign.

Differential Validity and Prediction: A Deeper Analysis of the New SAT *Paul Westrick, College Board; Jessica Marini, College Board; Emily Shaw, College Board*

This study extends differential validity and prediction research on the new SAT for student subgroups crossed by gender, race/ethnicity, and institutional admission selectivity. Results suggest that crossing subgroup categories provides valuable insights. These findings have particular implications for admission test criticisms that focus on issues of fairness.

From Theory to Practice: Applying Kane's Validity Framework to a Licensure Exam *Hong Qian, National Council of State Boards of Nursing*

Validity is the most fundamental consideration in developing tests and evaluating tests. However, few give it the attention it deserves. We describe our experience using Kane's validity framework to validate a large-scale high-stake licensure exam. Our research represents a work example others can follow when validating their assessments.

Grounding Assessment Development in Cultural Validity *Pohai Kukea Shultz, University of Hawaii; Kerry Englert, Seneca Consulting*

Historically, accountability assessments for Hawaiian language immersion schools have lacked cultural and linguistical sensitivity or possessed inadequate psychometric properties. The University of Hawaii and the Hawai'i Department of Education have partnered to develop a technically rigorous assessment that is grounded in the culture, language, and worldview of the Hawaiian community.

Chair:

Jeffrey T. Steedle, ACT

Discussant:

Chad Buckendahl, ACS Ventures, LLC

182. Methodological Considerations in IRT

Paper Session *2:15 to 3:45 pm*

Hilton San Francisco Union Square: Union Square 22

Participants:

A Study of Tests for Equal Rater Discrimination Across Rubric Categories *Xiaoliang Zhou, Columbia University; Lawrence DeCarlo, Columbia University*

When using rater models for constructed response scoring, it is assumed that raters discriminate equally well between different categories of the scoring rubric (it's related to a proportional odds assumption). Tests of this assumption are examined. Simulations show that the performance of a likelihood ratio test is excellent.

Effects of Extreme Response-Style Heterogeneity on the Estimated Correlations between Psychological Constructs *Daniel Adams, ETS; Dan Bolt, UW-Madison School of Education*

Response-style interference has the potential to bias measurements based on self-report rating-scale data. In this study we demonstrate through simulation that the degree in which extreme response style biases the estimated correlation between two scale scores is dependent on certain item- and person-level indicators.

Introduction to a longitudinal IRT model for measuring growth using accumulated test items *Minsung Kim, Buros Center for Testing; Hongwook Suh, Nebraska Department of Education; Kwanghee Jung, Texas Tech University*

A new longitudinal IRT model, accumulated longitudinal (AL) IRT model, was proposed to overcome several drawbacks of the existing models measuring students' growth in educational testing. A simulation study will be performed to evaluate its effectiveness on parameter recovery under general assessment conditions. Practical implications, limitations and further research directions will be discussed.

Investigation of Item Property Effects on Polytomous Items Using the Many-Facet Rasch Model *Jinho Kim, University of California, Berkeley; Mark Wilson, University of California, Berkeley; Jerred Jolin, University of California, Berkeley*

This paper discusses the applicability of item explanatory models using the Many-Facet Rasch Model for polytomous items, considering categorical item properties as facets. Two empirical studies demonstrated practical applications of those models to explain and predict overall item difficulties and showed methodological usefulness for items design and underlying hypothesis testing.

Precision-Weighted IRT Scale Transformations Via Response Function Methods *Alexander Weissman, Law School Admission Council; Wim J van der Linden, University of Twente*

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

Chair:

Francis O'Donnell, University of Massachusetts Amherst

Discussant:

Karen Draney, University of California Berkeley

183. Game-Based Assessment

Paper Session *2:15 to 3:45 pm*

Hilton San Francisco Union Square: Franciscan Ballroom A

Participants:

A Compensatory MIRT Model with Testlet Effects Applied in Game-based Assessment *Yao Xiong, Imbellus Inc.; Xinchu Zhao, Imbellus, INC; Marty McCall, Imbellus Inc.; Jack Buckley, Imbellus, Inc.*

A compensatory multidimensional item response theory (MIRT) model with testlet effects is proposed to account for testlet effects in multidimensional assessments. A simulation study is conducted to examine model performance under different testlet effect conditions. A real data collected from a game-based assessment is also analyzed to evaluate model applicability.

Developing Measures for Middle School Students' Data and Analysis Proficiencies Using A Game-Based Formative Assessment *Yuning Xu; Satabdi Basu, SRI International; Daisy Rutstein, SRI International*

This study developed measures for assessing middle school students' proficiencies in data and analysis, using log data from a game-based formative assessment. Results showed how the game can provide formative feedback that can be used by teachers. Assessment evidence identification was discussed under a principled assessment design framework.

Evidence for coachability resistance in stealth assessments *Christopher Stare, Imbellus, Inc.; Xinchu Zhao, Imbellus, Inc.; Alexander Thompson, Imbellus, Inc.; Erica Snow, Imbellus, Inc.; Jack Buckley, Imbellus, Inc.*

This study examined the coaching susceptibility of stealth assessments within a game-based environment. Participants were randomly assigned to either a control or a coached condition where they received coaching directly tied to the embedded stealth assessments. Results revealed no significant differences between performance in the control and coached groups.

Learning choices mediate the relation between prior academic achievement and performance *Maria Cutumisu, University of Alberta; Daniel Lewis Schwartz, Stanford University; Nigel Mantou Lou, University of Alberta*

Middle-school students (n=97) designed digital posters in a game-based assessment. Results showed that critical feedback-seeking fully mediated the link between students' prior academic achievement and their choice to revise their digital posters. Theoretical implications indicate that students' design-thinking strategies (critical feedback-seeking and revising) explain their performance, beyond prior academic achievement.

Modeling Collaborative Problem-Solving Skills from Team Interactions with an Educational Game *Maria Ofelia San Pedro, ACT, Inc.; Ruitao Liu, ACT*

Collaborative problem-solving components of cooperation, persistence and problem-solving were modeled from human-to-human team interactions in an educational game. Evidence-centered design was used to develop the tasks and indicators for each component, and Item Response Theory for scoring. Findings indicate behavior indicators may provide information not necessarily evident in self-reports.

Chair:

Molly Faulkner-Bond, WestEd

Discussant:

Susan Lottridge, AIR

184. Methodological Considerations #1

Paper Session *2:15 to 3:45 pm*

Hilton San Francisco Union Square: Franciscan Ballroom C

Participants:

Comparing Methods of Empirical Recovery of Mathematics Learning Trajectories Using Classroom Assessments *Emily Toutkoushian, North Carolina State University; Jere Confrey, North Carolina State University; Meetal Shah, North Carolina State University*

This study uses data from a middle-grades learning trajectory-aligned classroom assessment system to compare different methods and models of empirically recovering the trajectories within and across content areas. The results suggest how the methods can be modified for classroom assessments and which may be more appropriate for this data.

Comparing Six CCT Stopping Rules under Dual-Objective Adaptive Testing Using HIRT Models *Kuo-Feng Chang, University of Iowa; Won-Chan Lee, University of Iowa*

Recently, Chang and Lee (2019) proposed dual-objective adaptive testing, which simultaneously provides overall proficiencies as a summative assessment and subdomain ability classifications as a formative assessment for diagnostic purposes, using high-order IRT models. This study aims to compare the efficiency of six CCT stopping rules under their framework using simulation.

Detecting Test Speededness via Posterior Shift Using Responses and Response Times *Dmitry Belov, LSAC*

Modern detectors of speededness are based on assumption that speeded examinees perform increasingly worse as test progresses. However, this assumption may be often violated in practice due to various test taking behaviors. A new asymptotically optimal detector of speeded examinees is proposed that is not based on this assumption.

Improving the Efficiency of MCCT by Incorporating the Between-dimension Correlation Information *Bo Sien Hu, Institute of Education, National Sun Yat-sen University; Cheng Te Chen, Department of educational psychology and counseling, National Tsing Hua University; Ching Lin Shih, Institute of Education, National Sun Yat-sen University*

The efficiency of MCCT, that using ELR selection method and GLR stopping rule, was found mainly determined by the relative position of respondents' ability and cut-off points. A modified method that taking the between-dimension correlation into consideration was proposed in this study. Its performance was investigated through a simulation study.

Chair:

James McMillan, Virginia Commonwealth University

Discussant:

Tyler Holmes Matta, Pearson

174. Electronic Board. Monday 2:15

Electronic Board Session *2:15 to 3:45 pm*

Hilton San Francisco Union Square: Salon A

Participants:

A Logistic Regression Technique for Item Parameter Drift and Item Fit *Kevin Cappaert, Curriculum Associates; Brian F. Patterson, Curriculum Associates*

A two-step logistic regression (LR) technique for the detection of item parameter drift (IPD) and item misfit is proposed for use in a Rasch CAT environment. The proposed method compares LR derived difficulty and discrimination parameters from operational CAT responses and examinee estimates to pre-equated CAT item parameters.

An Empirical Example Comparison of Item Response Models Controlling Extreme Response Style *Huang Jianheng, City University of Hong Kong*

Four multidimensional item response models are developed to control extreme response style (ERS): ERS-GPCM, MNRM, IR tree model and UD tree model. This study compares these four models with GPCM through an empirical example. All four models can account for the ERS consistently. MNRM outperforms among other models.

An Investigation in UAMIRT on Testlet-Based Tests Equating under a NEAT Design *Qianqian Pan, The University of Hong Kong; Hongyu Diao, Educational Testing Service*

This study investigates the performance of unidimensional approximation of MIRT model on testlets equating under current calibration and separate calibration designs.

Assessing Nontrivial IRT Model Misfit in CAT *HWANGGYU LIM, Graduate Management Admission Council; Craig S. Wells, University of Massachusetts Amherst*

This study introduces a close-fit approach for assessing item response theory (IRT) model fit in computerized adaptive testing (CAT) using two item-fit statistics. The preliminary simulation results showed that the close-fit approach provided controlled Type I error rates and reasonable power for detecting items that have nontrivial model-data misfit.

Assessing the necessity of non-zero lower asymptote in the 3PL model application *Hirotaka Fukuhara, Pearson; Insu Paek, Florida State University*

In practice, the 3PLM is calibrated using a prior for the lower asymptote parameter (g). This makes traditional likelihood ratio test unsuable. This study examines the utlity of the several hypothesis test approaches based on the logit transformation of g to find a zero- or near zero lower asymptote item.

Capturing random guessers in tesltlet-based low-stakes assessments *Ying-Fang Chen, UC Berkeley; Hong Jiao, University of Maryland*

This study aims at simultaneously capturing testlet effects as well as test-taking strategy heterogeneity due to the presence of random guessers in low-stakes educational assessments within the Rasch measurement framework. Both simulation study and real data application are conducted to evaluate the proposed measurement model's performance.

Cross-classified Multilevel Item Response Theory Models with An Application to Higher Education Institutional Research *Sijia Huang, University of California, Los Angeles; Li Cai, University of California—Los Angeles; Alexandra Sturm, Loyola Marymount University; Abigail Panter, University of North Carolina at Chapel Hill; Viji Sathy, University of North Carolina at Chapel Hill*

In the present study, a cross-classified multilevel item response theory model is proposed. We introduce an efficient algorithm – the fully blocked Metropolis-Hastings Robbins-Monro algorithm – for maximum likelihood estimation. A simulation study is proposed to evaluate the performance of the algorithm. An empirical data from higher education is used for illustrations.

Detecting and Addressing Rasch Item Drift across Small and Moderate Samples *Jason P Kopp, American Board of Surgery; Andrew Jones, American Board of Surgery; Beatriz Ibanez, American Board of Surgery; Derek Sauder, James Madison University*

We compared six methods for detecting and addressing Rasch item drift with real and simulated data. Although some methods substantially improved examinee score bias, RMSE, and classification accuracy at large samples, improvements were reduced or nonexistent at small sample sizes.

Evaluating the Acceleration Model with Large-Scale Achievement Tests *Yuming Liu, ETS*

Simulated data based on empirical large-scale achievement tests are used to evaluate Samejima's (1997) acceleration model. Results of preliminary analyses show that the model is promising in terms of global goodness-of-fit statistics, robustness to prior distribution specifications, test length and sample size, and the accuracy of person parameter estimation.

Evaluation of the SAS IRT Procedure: Parameter Recovery and Item Fit *Nnamdi Ezike, University of Arkansas; Allison Ames, University of Arkansas; Brian C Leventhal, James Madison University*

SAS® item response theory (SAS-IRT) procedure is increasingly popular but not rigorously validated. This study uses simulation techniques to assess parameter recovery, Type I error, and power of SAS-IRT item fit functions. Practitioners are cautioned against Yen's Q_1 and G^2 fit indices when number of examinees and items are low.

Exploring a Novel Subscale-Level Multi-Stage Testing Approach for NAEP: A Simulation Study *Tong Wu, Purdue University; Young Yee Kim, American Institutes for Research; Xiaying (James) Zheng, American Institutes for Research*

In multi-stage testing, when the routing decision is made by the combined ability estimates from multiple subscales, the missingness in the next level is not at random, and item parameter estimates fit with multiple unidimensional IRT models are biased. This research explores if subscale level routing can overcome this issue.

Fitting Propensities of Item Response Models *Ezgi Ayturk, Fordham University; Leah Feuerstahler, Fordham University*

Functional form is an important yet often overlooked aspect of model complexity in the context of model comparison. This study explores the baseline tendencies of various unidimensional and multidimensional IRT models to show good fit to any given data as a function of their functional form through a simulation study.

Fixed Guessing 3PLM with Fixed Item Parameter Calibration without Estimating IRT-C Parameter *Sung-Hyuck Lee, GMAC; Kyung Han, GMAC*

The three-parameter logistic model (3PLM) has been widely used. However, the pseudo-guessing parameter is subject to unstable estimation even with a large sample. When the fixed guessing 3PLM (FG3PLM) is jointly used with the fixed item parameter calibration (FIPC), 3PLM still becomes a practically possible option for small scale programs.

Further Investigation of Anchor-Based Q-Sort Procedure for Judgmentally *Sharon Frey, Riverside Insights; David Swift, Riverside Insights; Emmett Cartwright, Riverside Insights; JP Kim, Riverside Insights.com*

The study demonstrates comparison of item difficulty parameters estimated (multiple regression and truncated mean) by a Q-sort methodology which uses expert ratings and known item parameters of embedded anchors (i.e., collect expert judgments of item difficulty without item field tryouts) with the parameters estimated via both field and operational testing.

Improving Efficiency of Multidimensional Item Response Model Adaptive Tests *Wenhao Wang, University of Kansas*

This study this study compared different CAT item selection and interim scoring methods of the three different confirmatory multidimensional IRT models with regard to efficiency and accuracy using simulation data. The results indicate simplified item selection and interim scoring method can provide accurate final scoring with small computation burden.

Independence of response time and accuracy: Analysis in massive CAT dataset *Ben Domingue, Stanford University; Klint Kanopka, Stanford University; Ben Stenhaug, Stanford University; James Soland, NWEA*

We study the assumption of conditional independence between response time and accuracy in a massive item response dataset from a large-scale standardized test. Results will offer generalizable insight on modeling of response time data. We also extend existing tests of this assumption to CAT settings.

Optimal Characteristics of Anchor Tests in Vertical Scaling *Gilbert Ngerano, UNCG/NBOME*

An investigation of empirical issues and complications that emerge when vertical scaling methods utilize nonequivalent groups with anchor test design to construct a vertical scale in order to better understand practical testing realities that would make vertical scaling to work or breakdown under different equating methods.

Semi-supervised learning method to adjust biased difficulty estimates caused by nonignorable missingness *Kang Xue, University of Georgia; A. Corinne Huggins-Manley, University of Florida; Walter Leite, University of Florida*

In data collected from virtual learning environments, nonignorable missing data patterns may impact the performance of applying psychometric models to item parameter and ability estimation. In this paper, we explored the factors related to missingness and adjusted the biased estimates using a semi-supervised learning method under 2PL-IRT.

Should Examinees Be Allowed to "Bank" Domain Scores? Evidence from Chiropractic Testing *Nai-En Tang, NBCE; Igor Himelfarb, National Board of Chiropractic Examiners; Bruce Shotts, NBCE*

Possible differences in parameter estimates between the first-time examinees who take the entire exam and repeaters who take only previously failed parts of the exam. Two exam administrations were selected for this study. Single-group and multi-group measurement invariance of the 2PL IRT model used to calibrate item responses was examined.

The Effect of Recalibrated Item Parameters on Estimated Scores *Jie Li, ACT, Inc.; Chunxin (Ann) Wang, ACT Inc; Yi He, ACT, Inc.*

Testing programs sometimes need to conduct item recalibration due to practical constraints encountered in testing operations. This study examines the impact of item recalibration on estimated scores and score distributions. Findings from the study will contribute to the development of valid, reliable and fair assessments for all.

185. NCME Board Meeting #2

Individual Submissions and Coordinated Sessions
Coordinated Session
4:00 to 7:00 pm
Hilton San Francisco Union Square: Golden Gate 1
NCME Board Meeting #2

187. **Invited Session Psychometricians without Borders: Expanding the Reach of Excellence in Measurement**

Coordinated Session *4:05 to 6:05 pm*

Hilton San Francisco Union Square: Plaza A

The NCME Mission Fund was established to provide a means for donors to express their tangible support for NCME's mission to advance the science and practice of measurement in education, and to provide individuals and organizations with financial support for projects, research, and travel that address this mission directly. Your generous donations provided the funding for the Mission Fund's first round of special initiatives designed to promote a broader understanding of high-quality assessment practices and appropriate test use among diverse groups of assessment stakeholders. The results of these initiatives will be presented and discussed.

Participants:

The current state of educational measurement and what we've learned *Kathryn N Thompson, James Madison University; Chi Hang Bryan Au, James Madison University; Brian C Leventhal, James Madison University*

Is it the Wrong Answer? Examining Stakeholder Voices in High Stakes Testing *Darius Taylor, University of Massachusetts, Amherst*

Increasing Measurement Literacy Through Social Media *Ren Liu, University of California, Merced; Ada Woo, ACTNext by ACT; A. Corinne Huggins-Manley, University of Florida*

Title: Improving Library Instruction Assessment Via Modern Measurement Practices *ZHEHAN JIANG, Baylor College of Medicine; Wenchao Ma, The University of Alabama*

Facilitating Assessment Knowledge: Parents-and-Teachers Tales *Maria Vasquez-Colina, Florida Atlantic University*

Session Organizer:

Michelle Boyer, Center for Assessment

Chair:

Michelle Boyer, Center for Assessment

188. Testing Time: The Push and Pull in High-Stakes State Accountability Assessments

Coordinated Session *4:05 to 6:05 pm*

Hilton San Francisco Union Square: Imperial Ballroom A

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

Presenter:

Rosie Read, Northgate High School, 2019 California Teacher of the Year,

Session Organizers:

Andrew Middlestead, Michigan Department of Education

Sue Brookhart, Dusquene University

Chair:

Vince Verges, Florida Department of Education

Discussants:

Joyce Zurkowski, Colorado Department of Education

Mark Reckase, Michigan State University

Kristen Huff, Curriculum Associates

189. Extending Automated Scoring Approaches Beyond Summary Scores: Leveraging Natural Language Processing to Support Educational Programs

Coordinated Session *4:05 to 6:05 pm*

Hilton San Francisco Union Square: Imperial Ballroom B

Traditionally, automated scoring techniques are used for modelling human hand scoring to score short or extended constructed response items on summative assessments for core subject areas like essay writing, English, math, history, reading, etc. This may lead to the impression that the most valuable use of these systems is to score examinee responses consistently, quickly, and cheaply. While that use of automated scoring systems the most widely deployed use of these systems, there are applications far beyond that use case! In this presentation, we will discuss several ways that these techniques are being used to support classrooms by fostering and assessing social skills like collaborative problem solving. Additionally, these automated scoring systems can power formative learning tools that not only assess students' answers to constructed response items but give students specific feedback about their points of confusion so that students can focus their future studies. Finally, these automated scoring techniques can support the very assessments and items they typical score. The non-summative scoring use cases above demonstrate that automated scoring techniques are equipped to positively support student learning and assessment by providing actionable information to several different areas of educational programs.

Participants:

Exploration of Automated Scoring Techniques for Measurement of Collaborative Problem- Solving Skills *Pravin Chopade, ACTNext by ACT; David Edwards, ACTNext by ACT; Alejandro Andrade, ACT Next; Saad Khan, ACTNext by ACT*

Using automated scoring to monitor and improve the assessment system *Jay Thomas, ACT, Inc.*

Automated Scoring & Feedback on Next Generation Science Standard Items *Gavin Henderson, ACT*

Session Organizer:

Gavin Henderson, ACT

Discussant:

Peter Foltz, Pearson

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

Coordinated Session *4:05 to 6:05 pm*

Hilton San Francisco Union Square: Yosemite A

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

Participants:

Using Latent Class Analysis to Explore the AI Score-ability of Constructed-Response Items *Kevin Raczynski, The University of Georgia; Hye-Jeong Choi, University of Georgia; Allan Cohen, University of Georgia*

Explaining Scores Produced from Neural Net-Based Engines *Susan Lottridge, AIR; Christopher M Ormerod, American Institutes for Research; Amir Jafari, AIR*

A Comparison of Latent Semantic Analysis and Latent Dirichlet Allocation *Jordan Wheeler, The University of Georgia; Allan Cohen, University of Georgia*

AES as an Aid in Quality Assurance of Essay Scoring *Yoav Cohen, National Institute for Testing & Evaluation, Israel; Effi Levi, The National Institute for Testing & Evaluation*

Session Organizer:

Kevin Raczynski, The University of Georgia

Discussant:

Mark David Shermis, American University of Bahrain

191. Psychometric Innovations and Advances in Medical Educational Assessments

Coordinated Session *4:05 to 6:05 pm*

Hilton San Francisco Union Square: Yosemite B

The proposed coordinated session provides an overview of new and innovative psychometric approaches based on data science methods and item response theory (IRT) model extensions for improving the quality, validity, and accuracy of test items, test designs and test scores in the context of medical licensing exams and medical education. In the assessment of cognitive constructs, methods such as natural language processing (NLP), machine learning, feature generation, and cluster analyses of process data provide additional information about examinee ability and item characteristics. This information can be used to improve the test development process and the efficiency of test designs, allow for the automated coding and scoring of open-ended or constructed responses for an increased measurement precision, and a better interpretation of generated process data features in relation to test scores. Moreover, new IRT model extensions allow the modeling of different latent variables related to response biases and measurement error for a more valid interpretation of noncognitive constructs. These new approaches and model extensions are not only helpful for the enhancement of high stakes tests, but can also be used to advance the education and professional development of medical students, young physicians and practitioners prior to and beyond licensure testing.

Participants:

Helping Item Writers by Making NLP-Based Suggestions for Item Distractors *Peter Baldwin, National Board of Medical Examiners (NBME); Victoria Yaneva, National Board of Medical Examiners; Janet Mee, National Board of Medical Examiners; Brian Clauser, National Board of Medical Examiners; Le An Ha, University of Wolverhampton*

On the Utility of Using Transfer Learning to Predict Item Characteristics *Kang Xue, University of Georgia; Victoria Yaneva, National Board of Medical Examiners; Christopher Runyon, NBME*

Exploring Automated Assessment of Spoken English Proficiency for Medical Licensure Exams *Su G. Somay, National Board of Medical Examiners (NBME); Jessica Salt, Education Commission for Foreign Medican Graduates (ECFMG)*

An Automated Scoring Routine for Constructed Responses on a Medical Licensure Exam *Christopher Runyon, NBME; Polina Harik, National Board of Medical Examiners; Abeed Sarker, Emory University School of Medicine; Graciela Gonzalex-Hernandex, University of Pennsylvania*

Identifying Response Pattern in Clinical Notes with NLP Feature Identification *Janet Mee, National Board of Medical Examiners; Ravi Pandian, National Board of Medical Examiners; Andrew Houriet, National Board of Medical Examiners; Christopher Yang, Drexel University*

IRTTree Response Style Modeling for Improving Feedbacks about Cognitive Biases *Lale Khorramdel, National Board of Medical Examiners; Matthias von Davier, National Board of Medical Examiners; Ann King, National Board of Medical Examiners; Andrew Houriet, National Board of Medical Examiners*

Session Organizers:

Lale Khorramdel, National Board of Medical Examiners
Christopher Runyon, NBME

Chair:

Lale Khorramdel, National Board of Medical Examiners

Discussant:

Isaac I Bejar, Educational Testing Service

192. Computerized Adaptive Testing, AI and Smart Learning

Coordinated Session *4:05 to 6:05 pm*

Hilton San Francisco Union Square: Yosemite C

With current big data, AI and deep learning uprising, psychometricians are at a crossroads whether to make a turn to the fashionable machine learning and deep learning. Consisting of five presentations, this coordinated session is to show that computerized adaptive testing (CAT), which is considered as one of the earliest AI applications in Educational Testing, will continue to play an important role in both measurement and learning in the big data era. CAT has been well advanced during the last 40 years. The rapid developments in technology have made large-scale CAT implementations easier than ever before. However, CAT has not been well acknowledged by data scientists. Today we will highlight our discussion on whether CAT can help AI in educational research, in particular, how to incorporate "deep learning" to support smart testing and smart teaching. Our goal is to build many reliable, and also affordable, web-based diagnostic tools for schools to automatically classify students' mastery levels for any given set of cognitive skills that students need to master. In addition, we will show how the tools can be employed to support individualized learning on a mass scale.

Participants:

1. AI and Machine Learning in Psychometrics? Old News. *Nathan Thompson, Assessment Systems Corporation*
2. Adaptive Learning System Design with Deep Reinforcement Learning and Neural Networks *Xiao Li, University of Illinois at Urbana-Champaign; Jinming Zhang, University of Illinois at Urbana-Champaign; Hua-Hua Chang, Purdue University*
3. Data-drive Attribute Hierarchy Detection Using Bayesian Graphical Model *Yinghan Chen, University of Nevada Reno; Shiyu Wang, University of Georgia*
4. On-The-Fly Parameter Estimation Based on Item Response Theory in Adaptive Learning Systems *Shengyu Jiang, University of Minnesoda; Chun Wang, University of Washington*
5. Improving Scoring Precision with Features Extracted from Log Data *Susu Zhang, Columbia University; Xueying Tang, University of Arizona; Jingchen Liu, Columbia University; Qiwei He, Educational Testing Service*

Session Organizer:

Hua-Hua Hua Chang, Purdue University

Chair:

Hua-Hua Hua Chang, Purdue University

Discussant:

Hua-Hua Hua Chang, Purdue University

193. Predictive Standard Setting: Improving the Method, Debating the Madness

Coordinated Session – Panel Discussion *4:05 to 6:05 pm*

Hilton San Francisco Union Square: Plaza B

Test scores measure, and test scores predict. Predictions can anchor statements about current performance in terms of future outcomes—including test scores, grades, and graduation—through a process called "predictive standard setting." Presenters in this symposium will debate how and whether predictions should inform standard setting, whether standards should make predictions, and how predictions should count as validity evidence. Contexts include the SAT and ACT college readiness benchmarks, state accountability tests in grades 3-8, interim assessments, and NAEP. These issues are salient as educational policies, policymakers, and practitioners value these predictions in "career and college readiness" frameworks. Presenters will discuss and debate advances in three stages of predictive standard setting: 1) generating accurate predictive statements using statistical methods; 2) managing predictive data in the standard setting process; then 3) communicating results using benchmark and achievement-level descriptors. Some presenters believe strongly that predictive statements build valid consensus among standard setting panelists and help users understand the meaning and relevance of scores. Other presenters believe strongly that predictive statements build false consensus and subjugate the subject-matter relevance of scores in favor of ambiguous future outcomes. Presenters will give short presentations and then engage in moderated discussion with each other and the audience.

Presenters:

Jennifer Beimers, Pearson
Wayne J. Camara, ACT
Laurie Laughlin Davis, Curriculum Associates
Laura Hamilton, RAND
Deanna Morgan, College Board
Yi Xe Thng, Singapore Ministry of Education

Session Organizer:

Andrew Ho, Harvard Graduate School of Education

Chair:

Walter (Denny) Way, College Board

194. The Rating of Writing Skills in Secondary Education in Europe

Coordinated Session *4:05 to 6:05 pm*

Hilton San Francisco Union Square: Franciscan Ballroom A

The symposium presents research on the rating of writing proficiency in both first and second language in Europe. So far, there have been few comprehensive investigations of this aspect of language proficiency in the context of educational assessment in Europe. This is surprising considering the practical importance of writing in the context of high- and low-stakes assessment in secondary and tertiary education. To conduct empirical research on writing skills, there is a need to generate valid ratings regarding the quality of students' texts. First, we present research on standardized large-scale assessments, drawing on judgments of professionally trained raters and automated essay scoring. Second, we consider the school perspective by including samples of text assessments by teachers with different levels of experience. Third, we present cutting-edge research on automated assessment of writing skills. Three methodological advances and their applications are presented and discussed: 1) a novel multilevel modeling approach to investigate predictors of judgment accuracy, 2) different automated scoring approaches and 3) machine learning techniques. In the different papers, we explore synergies between these approaches. Overall, the session aims at giving its audience a comprehensive overview of cutting-edge research perspectives on conceptualizing and measuring writing skills in the European context.

Participants:

Context and Overview *Stefan Keller, FHNW*

Effects of Teachers' Characteristics on Judgment Accuracy: a Multilevel Modeling Approach *Johanna Fleckenstein, IPN Kiel; Steffen Zitzmann, IPL Uni Kiel; Thorben Jansen, IPL Uni Kiel; Jennifer Meyer, IPN Kiel; Jens Möller, IPL Uni Kiel*

Effects of Teachers' Characteristics on Judgment Accuracy in L1 Writing *Thorben Jansen, IPL Uni Kiel; Raja Reble, University of Kiel; Jens Möller, IPL Uni Kiel*

Applying Machine Learning in the European Context: Linear Regression versus Boosting Approaches *Jennifer Meyer, IPN Kiel; Thorben Jansen, IPL Uni Kiel; Johanna Fleckenstein, IPN Kiel; Olaf Köller, IPN Kiel*

Language Technology Support for Analyzing Learner Texts *Torsten Zesch, University of Duisburg-Essen; Andrea Horbach, University of Duisburg-Essen; Ronja Laarmann-Quante, University of Bochum*

Session Organizer:

Jennifer Meyer, IPN Kiel

Chair:

Olaf Köller, IPN Kiel

Discussants:

Andre Alexander Rupp, Educational Testing Service (ETS)

Paul Deane, ETS

195. Research Blitz - Fairness

Research Blitz Session *4:05 to 6:05 pm*

Hilton San Francisco Union Square: Union Square 19/20

Participants:

A Meta-Analysis on the Effectiveness of English Learner Testing Accommodations *Samuel D Ihlenfeldt, University of Minnesota; Joseph A. Rios, University of Minnesota; Carlos Chavez, University of Minnesota*

Addressing a gap in scholarship on the effectiveness of testing accommodations for EL students, this meta-analysis of 26 studies and 95 effect sizes (N = 11,069) shows that accommodations improved test performance by 0.16 standard deviations. However, individual accommodation effects were not statistically significant, suggesting a need for further research.

An approach to developing and evaluating equitable, culturally-sensitive large-scale assessment systems *Stanley N Rabinowitz, Pearson; Edynn Sato, Sato Education Consulting LLC*

We describe how cultural-based learning orientations and linguistic-based meaning-making that differs across linguistically and culturally diverse students impact assessment outcomes and the validity of interpretations of these outcomes, as well as implications for extending our current principled approaches to assessment design and development for our diverse learners.

Efficacy of Balanced Literacy Instruction on Students' Reading Performances: Using PIRLS Data *lu guo, Texas Tech University; Jian Wang, Texas Tech University*

Abstract This study investigated on the overall effects of balanced literacy (BL) instruction and influence of each component of BL instruction on English Language Learners (ELLs) and native-speaking students' reading outcomes. Keywords: Balanced Literacy, PIRLS, ELLs

Game-based Spoken Interaction Assessment in Special Need Children *Jos Keuning, Cito*

Staying on topic, aligning to interlocutors, using non-verbal communication: skills that are all needed in a conversation. How can you assess these skills as a teacher? And how do you do that in children attending special education? This study addresses these questions using a collaborative board game: the so-called Fischerspiel.

Generating Individual Difference Profiles via Cluster Analysis: Toward Caring Assessments for Science *Jesse R. Sparks, Educational Testing Service; Jonathan Steinberg, Educational Testing Service; Karen Castellano, Educational Testing Service; Blair Lehman, ETS; Diego Zapata, ETS*

Individual differences in students' personal qualities (e.g., growth mindset) have significant relationships to performance on assessment tasks. Cluster analyses yielded four distinct profiles of middle school students (N=630), based on measures of grit, cognitive flexibility, growth mindset, and test anxiety. Science assessment performance significantly differed based on students' profile membership.

Latent Dirichlet Analysis to Aid Equity in Identification for Gifted Education *Holmes Finch, Ball State University; Maria Finch, Ball State University; Brian F French, Washington State University; Claire Braun, Ball State University*

An innovative assessment was developed to increase fairness and accuracy in gifted education identification. Topic-modeling via latent dirichlet analysis was leveraged to extract themes from 13 written responses by parents. Scores differed between student groups as expected, and correlations with other measures were in expected direction and magnitude.

Making Assessment Matter for Diverse Students—Promoting Culturally Responsive Assessments *Guri A. Nortvedt, University of Oslo*

In assessment situations, all students ideally have the same opportunities to demonstrate their competence. However, large differences in favor of majority students are often seen. This paper reports a review study aiming to identify the obstacles to a culturally fair assessment, as well as the components comprising culturally responsive assessment.

Measurable Support: Applying Polytomous Scoring to Single-Select Multiple-Choice Items with Scaffolding *Elena Nightingale, Georgia Department of Education; Jan Reyes, Georgia Department of Education*

This mixed-methods exploratory analysis of scaffolding function in the inaugural administration of the Georgia Alternate Assessment (GAA) 2.0 is guided by themes identified in educator feedback, answering the question: Does scaffolding provide meaningful (measurable/scorable) additional information about student performance? Complexities such as repeated selection and varied scaffolding types are presented.

Middle School Students' Mathematics Opportunity-to-learn and Its Impact on the Mathematics Achievement *Gao Ruiyan, The University of Hong Kong; Frederick Leung, The University of Hong Kong; Tao Yang, Beijing Normal University*

The study investigated the mediation mechanisms according to the opportunity-propensity framework. In Hong Kong, the mediation path of gender, ESCS - opportunity-to-learn - mathematics self-efficacy - mathematics achievement is significant; while in Shanghai, the mediation path of age, grade - opportunity-to-learn - mathematics interest - mathematics achievement is significant.

Understanding Fairness *Lissette Tolentino, University of Florida; A. Corinne Huggins-Manley, University of Florida*

Understanding how teachers comprehend fairness regarding standardized testing is important to the educational measurement community. The purpose of this study is to understand teachers' perceptions of fairness in relation to marginalized communities. This research aims to provide insights into the communication and perceptions of fairness amongst stakeholders and test users.

Using propensity score matching to understand the impacts of test accommodation on subgroups *Sarah Tran, NWEA; Wei He, NWEA*

This study intends to examine the effects of test accommodation with propensity score matching under the differential boost framework, using a large sample of Grades 3 to 8 reading tests from a large-scale K-12 assessment administered in the form of computerized adaptive test.

Why Can't I Just Translate? Transadaptation of a SEL Assessment in Brazil *Cristina Anguiano-Carrasco, ACT, Inc.; Caio Lo Binco, LIV; Dana Murano, ACT; Joana London, LIV; Jeremy Burrus, ACT, Inc*

Global interest in social and emotional learning has exploded in the past years but there are few psychometrically sound SEL assessments available outside of the US. This paper discusses the development and application of an ITC-guided transadaptation process to transadapt a US-based assessment for use in Brazil.

Chair:

Kristin Morrison, Curriculum Associates

196. Research Blitz - Reliability

Research Blitz Session *4:05 to 6:05 pm*

Hilton San Francisco Union Square: Union Square 15/16

Participants:

Classification Accuracy in High-Stakes Exam: CTT, IRT and Bayesian Methods *Igor Himelfarb, National Board of Chiropractic Examiners; Bruce Shotts, NBCE*

This study was conducted to investigate classification accuracy between the full and reduce forms of the Chiropractic Clinical Sciences Exam (Part II) using methodologies based on classical test theory (CTT), item response theory (IRT) and Bayes' Theorem. Results showed high accuracy consistent across all three methods.

Comparing Reliability Methods for Diagnostic Mastery Classifications *William Thompson, University of Kansas; Brooke Nash, University of Kansas; Amy K Clark, University of Kansas*

Although many methods have been proposed for measuring the reliability of diagnostic mastery classifications, these methods have not yet been directly compared in an applied operational setting. In this paper, three popular methods for estimating this are investigated an operational assessment using diagnostic classification models.

Estimating Classification Decisions for Incomplete Tests *Richard Feinberg, National Board of Medical Examiners*

Interruptions during administration of large-scale testing programs are inevitable. For situations where the primary purpose is classification, such as a pass/fail decision, the current study explores three methods to estimate that decision based on a partially completed testing attempt. Applications and comparisons using large-scale operational data are presented.

Estimation of Multilevel Reliability for Social and Emotional Learning Measures *Carlos Chavez, University of Minnesota; Michael Clifford Rodriguez, University of Minnesota; Kyle Nickodem, University of Minnesota; Kory Vue, University of Minnesota; Tai Do, University of Minnesota*

Although there is growing interest in using social emotional learning measures for school accountability, there is a dearth of evidence to support the use of these measures at the school level. This study provides evidence for the reliability of SEL measures at the school level.

Evaluating mastery of sub-domains *Anton Beguin, Cito; Hendrik Straat, Cito*

This study investigates the way in measurement of proficiency and evaluation of mastery of (sub-)domains can be combined. Mastery is evaluated using informative Bayesian hypotheses and both linear and adaptive designs are discussed.

Evaluation of Subscore Reporting for English Language Assessments of Young Students *Tongyun Li, Educational Testing Service; Jiyun Zu, Educational Testing Service*

This study is an investigation of subscore reporting in two large-scale English assessments of young students. CTT-based approaches proposed by Haberman (2005) are used to evaluate whether it is meaningful to report subscores at individual and institution levels. Results illustrate institution-level subscore reporting is justifiable given a reasonable sample size.

How Days Between Testing Impacts Alternate Forms Reliability in Computerized Adaptive Tests *Adam E Wyse, Renaissance*

This study uses data from math and reading computerized adaptive tests to explore how the number of days between tests impacts alternate forms reliability coefficients. Results suggest that alternate forms reliability coefficients were lower when the second test was administered within three to four weeks of the first test.

Hypothesis Testing for Independent and Dependent Coefficient Alpha Reliability *Rashid S Almehrizi, Sultan Qaboos University; Rehab Alhakmani, Ministry of Education*

It is expected that assessment tools maintain its reliability across demographic variables and over repeated times. Hypothesis testing of equality of population coefficient alpha over independent and dependent administrations of tests is used. The paper presents asymptotic estimates of sampling variance of coefficient alpha and the sampling covariance of two paired alpha coefficients using delta method. An independent samples z-test and a paired samples z-test are implemented.

Understanding Different Ways to Compute Measurement Errors and Score Reliability for Adaptive Tests *yiqin pan, University of Wisconsin-Madison; Sung-Hyuck Lee, GMAC; Kyung Han, GMAC*

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

Use of Assessment Results in Presence of Model Misspecification and Measurement Error *Salih Binici, Florida Department of Education; Yachen Luo, Florida State University*

This study examines consequences of model misfit and measurement error on reporting outcomes for a large-scale assessment. It investigates whether ignoring model misspecification and measurement error has any practical impact on reported scale scores for parents and teachers, also their secondary use in statistical analyses to inform policy makers.

Chair:

Kimberly Colvin, University at Albany, SUNY

197. Research Blitz - Response Time

Research Blitz Session *4:05 to 6:05 pm*

Hilton San Francisco Union Square: Union Square 22

Participants:

Detecting Examinees with Aberrant Behaviors Using Item Response Times and Person-Fit *NOOREE HUH, ACT, Inc.; Chi-Yu Huang, ACT, Inc.; Yang Lu, ACT, Inc.*

This study will evaluate the potential usefulness of the combination of item response times and person-fit methods in detecting possible cheaters in online testing. Data will be simulated by considering the examinees' ability levels, the number of breached items, the difficulty of breached items, and the test lengths.

Differential Speededness on the MCAT® Exam *Ying Jin, Association of American Medical Colleges; Jordan Prendez, HumRRO; Marc Kroopnick, Association of American Medical Colleges*

This study examined the extent to which the Medical College Admission Test (MCAT) is differentially speeded for Black and Hispanic examinees relative to white examinees. The findings suggest that the time constraints on the MCAT exam affect examinee performance equally for Black and Hispanic examinees relative to white examinees.

Exploring Differential Item Time Functioning through Jointly Modeling Ability and Response Time *Young Yee Kim, American Institutes for Research; Nixi Wang, University of Washington; Xiaying (James) Zheng, American Institutes for Research; XIAOYING FENG, Avar Consulting, Inc; American Institutes for Research (contractor); Markus Broer, American Institutes for Research*

We define items that appear to be more or less time intensive for certain subgroups, conditioning on their ability and speed, as exhibiting differential item time functioning (DITF). We explore DITF by using a confirmatory factor analysis approach to test measurement invariance with joint modeling of response and response time.

Growth Mixture Response Time Model for a Longitudinal Medical Certification Exam *Luping Niu, The University of Texas at Austin; Drew Dallas, NCCPA*

The present study proposed a model for a longitudinal medical exam to investigate differentiated response time(RT) patterns, the growth of person's speed, how speed growth differs across persons, and the source (e.g., guessing, learning, specialty, adapting procedure, etc.) of variability of the speed progress and RT patterns.

Identify Rapid-Guessing Behavior for a Working-Memory Test Using the Lognormal Model *Ping Yin, HumRRO; Mary Pommerich, Defense Manpower Data Center*

In a computer adaptive testing environment, both the item response and response time can be modeled to evaluate person fit. This study applies the lognormal response time model to a working-memory test and evaluates whether rapid-guessing behavior can be identified using a person-fit statistic.

Is Speed Really Constant? Investigating Test-taker Pacing Behaviors *Tanesia Beverly, Law School Admission Council; Alexander Weissman, Law School Admission Council*

This study test-takers' pacing behaviors using a generalized common factor model for variable speed and ability. We found three distinct phases of pacing behavior within a test. Test takers may exhibit within each phase constant, increasing, or decreasing speed.

Prediction of Testing Time with the Decision Tree Model *Yang Lu, ACT, Inc.; Ruitao Liu, ACT; Yu Fang, Law School Admission Council; Lu Wang, ACT*

This paper is to develop and evaluate a new method to predict test response time based on the decision tree model. The results will be compared with predicted response time from the traditional hierarchical linear model and observed response time from the real data that show no speededness issues.

The Effects of Time-Limits on Computer Adaptive Test Scores *Justin L. Kern, University of Illinois at Urbana-Champaign*

It has been argued that time-limits have an unintended effect on the validity of test scores which must be studied to a greater extent (Lu & Sireci, 2007). This study will look at the effects of time-limits on test scores in a computer adaptive test (CAT) situation.

Using Item Response Times to Screen for At-risk Students: Reading Comprehension *Mark L Davison, University of Minnesota; Bowen Liu, University of Minnesota; Patrick C Kennedy, University of Oregon; Sarah E Carlson, Georgia State University; Ben Seipel, California State University, Chico; Gina Biancarosa, University of Oregon*

Item response times for 3rd, 4th, and 5th graders on a multiple-choice, computer administered reading comprehension test were used to predict which students achieved proficiency on a statewide reading test. Response times improved prediction, over and above response accuracy, in all three grades, more so in 4th and 5th.

Using Test Timing Data to Provide Insight into Test Taker Performance *Matthew T Schultz, AICPA; Joshua I Stopek, AICPA*

Implementing principled assessment frameworks (PAFs) impacts the way that subject matter experts think about the nature of their construct. This paper focuses on test specifications derived from a PAF approach, specifically focusing on the relationship between skill level, item features and timing. Implications for validity evidence and operational implementation are considered.

Chair:

Deborah Schnipke, ACS Ventures, LLC

186. Electronic Board. Monday 4:05

Electronic Board Session *4:05 to 6:05 pm*

Hilton San Francisco Union Square: Salon A

Participants:

A Simulation Based Approach for Evaluating SEM Models with Many Observed Variables *Dexin Shi, University of South Carolina; Taehun Lee, Chung-Ang University; Amanda Fairchild, University of South Carolina; ZHEHAN JIANG, Baylor College of Medicine*

Fitting a large structural equation modeling model with moderate to small sample sizes results in inflated Type I errors for the likelihood ratio test statistic. We introduce a simulation-based approach to adjust for Type 1 error inflation in the model $\chi 2$ for testing model fit in large SEM models.

A Statistical Method for Q-Matrix Specification in Cognitive Diagnosis Models *Uk Hyun Cho, University of North Carolina at Greensboro; Kun Su, UNCG; Robert Henson, University of North Carolina at Greensboro*

This research explores the use of a statistical method to specify Q-matrix for five different Cognitive Diagnosis Models. The five models represent the range from the conjunctive to disjunctive continuum. The relationships between the performance of the statistical method, the model selection and the type of Q-matrix are discussed.

A Supervised Topic Model for Analyzing Answers in Constructed Response Tests *Hye-Jeong Choi, University of Georgia; Seohyun Kim, University of Georgia; Jonathan Templin, University of Iowa; Allan Cohen, University of Georgia*

In this study, we present a supervised topic model for analyzing text in an educational test. The motivation of this study is to better describe examinees' knowledge by combining the strengths of a topic model and a diagnostic model for analyzing simultaneously accuracy responses and text data.

Accurately Estimating Ability Levels of Examinees who Copy Answers *Sarah L Toton, Caveon; Dennis Maynes, Caveon*

Examinees who copy answers have invalid ability estimates. We propose using an initial estimate of the rate of answer-copying in a mixture model to obtain refined estimates of copier ability. This research will allow more accurate ability estimation and will remove a source of bias in estimating answer-copying rates.

Achievement Gaps for the ECLS-K 2011 Assessments Using a NAEP-like Conditioning Model *Soo Lee, American Institutes for Research; Burhan Ogut, American Institutes for Research; Markus Broer, American Institutes for Research*

The conditioning model refers to a process that incorporates both cognitive item responses as well as student's additional background information (e.g., student socioeconomic status; SES). This study aimed to investigate achievement gaps for ECLS-K: 2011 assessments if a NAEP-like "conditioning model" is applied for some major student subgroups.

An Empirical Study of Omitted Answers on a Language Proficiency Test *Merve Sarac, University of Wisconsin-Madison; Eric Loken, University of Connecticut*

For number-right scored tests, the optimal strategy is not to omit. Yet, on a large-scale test of English as a second language, we found a sizable proportion of omits – most frequently on reading passages. Further investigation revealed that non-omitting test-takers were guessing at a high rate on reading passages.

Applying Natural Language Processing and Deep Learning to Better Understand Test Wiseness *Chris Foster, Caveon*

Test wiseness allows examinees to increase their probability of getting a correct response to an item without knowing the item content. In this paper we discuss strategies using modern machine learning natural language processing techniques to build a model which can get items correct without content knowledge.

Attribute Hierarchy Models in Cognitive Diagnosis: Conditions of Q-Matrix Completeness *Hans Friedrich Koehn, Department of Psychology, University of Illinois at Urbana-Champaign*

Attribute Hierarchy Models in cognitive diagnosis account for dependencies among attributes by imposing prerequisite relations on attribute mastery. Thus, constructing a complete Q-matrix may be difficult because many attribute combinations are no longer admissible because they violate the prerequisite structure; hence, certain item attribute profiles are not meaningfully defined.

Contrasting Groups Approach: Setting Multiple Cut Scores for a Complex Performance Examination *Fang Tian, Medical Council of Canada; Andrea Gotzmann, Medical Council of Canada; Sirius Qin, Medical Council of Canada; Maxim Morin, Medical Council of Canada; Liane Patsula, Medical Council of Canada; Andre De Champlain, Medical Council of Canada*

We conducted a Generalizability analysis of the judgment ratings collected using the Contrasting Groups method for setting two cut scores for a complex clinical performance examination. The findings support our multilevel application of the Contrasting Groups method and the generalizability of the standard-setting results.

Does Prior Exposure to Test Form/Items Affect an Examinee's Performance? *Avi Allalouf, NITE; Marina Fronton, NITE; Tony Gutentag, Hebrew University, Jerusalem*

The effect of prior exposure to test form/items was investigated using data on 13,183 examinees in English, Quantitative and Verbal reasoning. Examinees mostly do not change their answer in the second administration; only small improvements were noted. There was no systematic relation between repetition effects and time lapse between administrations.

Does Splitting a Test into Multiple Sessions Reduce Rapid Guessing? *Audra Kosh, Edmentum*

In low-stakes formative testing programs, teachers often maintain control over whether to administer a test in one lengthy session or divide the test into multiple shorter sessions. This study examines how each of these administration conditions relates to the frequency of rapid guessing in K-8 math and reading tests.

Effectiveness of DupER Augmentation in Improving IRT-3PL Calibrations with Small Samples *Guanlan Xu, University of Iowa; Walter P. Vispoel, University of Iowa*

We investigated the effectiveness of Duplicate, Erase, and Replace Augmentation procedures (DupER; Foley, 2010) in calibrating the 3PL model using small datasets in an operational setting. DupER procedures were most effective with medium-sized samples (n = 600) in which imputed datasets adequately reflected the score distribution within the target population.

Examining the Impact of Students' Engagement Levels on Their Survey Scores *Munevver Ilgun Dibek, TED University; H. Cigdem Yavuz,*

Cukurova University

This study aims to examine how engagement levels of the students participating in PISA 2015 play a role on their survey scores by comparing two different response time measures. The students' engagement level indexes significantly predicted survey scores while the index suggested by Huang et al. (2012) showed better fit.

Examining the Performance of Alignment Method in DIF Analyses *Paulius Satkus, James Madison University; Christine E DeMars, James Madison University*

The alignment procedure (Muthén and Asparouhov (2014) offers several advantages over the traditional methods in differential item functioning (DIF) analyses. In a simulation study, we evaluated how alignment adjusts the estimated group means given different DIF patterns. Results suggests that the alignment method is not robust to several DIF patterns.

Investigating Attribute Hierarchical Relations with Multilevel Mediation Measurement Modeling *Yi-Hsin Chen, University of South Florida; Zhiyao Yi, University of South Florida; Denisse Thompson, University of South Florida*

This study applied multilevel structural equation modeling (MLSEM) to investigate attribute hierarchical relations among four Geometry thinking skills based on the van Hiele theory. The results indicated that the completed mediation model was the best fit model, meaning that four skills have a linearly sequential and hierarchical relation.

Investigating two methods of detecting group-level potential cheating on standardized exams *Myung Hee Im, Americian Institutes for Research; CHIH-KAI LIN, American Institutes for Research; Yuan Hong, Americian Institutes for Research*

Due to increased cheating on standardized exams that can threaten the security and validity of score interpretation, the present simulation study evaluates two methods of detecting unusual test score changes between years and aberrant item response patterns, especially focusing on the group-level abnormality.

Joint Analysis of Social and Item Response Networks with Latent Space Models *Shuo Wang, The Ohio State University; Subhadeep Paul, The Ohio State University; Jessica Logan, The Ohio State University; Paul De Boeck, Ohio State University*

We propose a latent space model for heterogeneous networks (LSMH) to jointly analyze social network and item response data by combining the network analysis with the item response theory. Using the LSMH, researchers can study the connection between friendship and school adjustment and identify students with difficulties adjusting to school.

Scaling Rater Parameter Estimates from Rater Response Models under the NEAT Design *Jodi Casabianca, Educational Testing Service; John Donoghue, Educational Testing Service; Szu-Fu Chao, Educational Testing Service*

Rater response models, which quantify individual rater effects, can be leveraged in large-scale testing situations in which there are several hundred raters scoring multiple prompts on a regular basis. This paper discusses results of simulation/empirical analyses focused on the challenges of scaling estimates from these models for rater monitoring.

Score-level Sample Size Requirements for Technology-enhanced Items: A Simulation Study *Shu-chuan Kao, Pearson; William Muntean, Pearson; Joe Betts, NCSBN*

The use of technology-enhanced items (TEIs) brings great possibilities for item development and item scoring. This study explores the impact of insufficient score-level sample size on item characteristics and data-model for TEIs that can be better interpreted by polytomous scoring with the use of partial credit model.

What do Students Know about Forces and Motion?: An Application of Cognitive Diagnostic Models *Dongsheng Dong, University of Washington; Min Li, University of Washington; Klint Kanopka, Stanford University; Philip Hernandez, Stanford University; Jim Minstrell, FACET Innovations; Maria Araceli Ruiz-Primo, Stanford University*

This paper examines students' understanding of key physics concepts around forces and motion in a middle-school physics test. The goals of this paper are to illustrate how to capture students' thinking using a pre-defined attribute matrix and cognitive diagnostic models, taking into account the local dependence between two-tiered items.

198. Diversity Committee reception

6:30 to 8:00 pm

Hilton San Francisco Union Square: Salon B

Diversity Committee reception

Session Organizers:

Leanne R. Ketterlin-Geller, Southern Methodist University