NCME

national council on measurement in education

APRIL 4-8, 2019

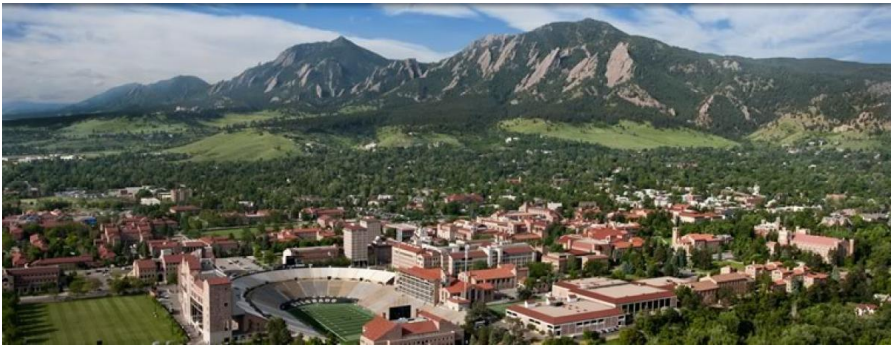
Fairmont Royal York, Toronto, Canada

NCME

THIRD ANNUAL CONFERENCE SPECIAL CONFERENCE ON CLASSROOM ASSESSMENT

September 18 – 19, 2019

"CLASSROOM ASSESSMENT AS A LEARNING EXPERIENCE"



JOIN US for the 3rd National Council on Measurement in Education (NCME) Special Conference on Classroom Assessment hosted by CADRE and CU Boulder's School of Education in partnership with:

- Aurora Public Schools
- Cherry Creek School District
- Colorado Department of Education
- Denver Public Schools
- National Center for the Improvement of Educational Assessment

REGISTER HERE:
https://www.colorado.edu/cadre/classroom-assessment-learning-experience

WELCOME FROM THE PROGRAM CHAIRS

Welcome to Toronto! We're so happy you're joining us for the 2019 NCME Annual Meeting. We've collaborated with many of you to prepare a program that we hope offers opportunities to learn, grow, connect, and celebrate some incredible work and achievements in the field.

This year's conference theme of "Communicating with the Public about Educational Measurement" allowed us to place some well-deserved focus within the program on how to more broadly communicate with the world about our work, how to present our work in clearer and more compelling ways, and how to reach diverse audiences. We still have many technical sessions, but we also have invited sessions such as ***Communicating/Depicting Results in Easily Accessible Ways, Across Broad Audiences*** (Saturday, 12:20pm); ***Appropriately Interpreting, Comparing, and Communicating Results from International Assessments: Challenges and Opportunities*** (Monday, 8am); and ***The Influence of Stakeholder Needs and Values on Assessment Design and Reporting*** (Sunday, 3:20pm). Of particular note, NCME President Rebecca Zwick has organized a special session for attendees, ***Communicating Your Research to the Media***, with Emily Richmond (public editor of the Education Writers Association) and Holly Yettick (director of the Education Week Research Center), who will share the perspective of editors and journalists combing through various studies and press releases to determine what is newsworthy and describe the process from pitch to published story. This session will be on Saturday at 10:25am. Another timely and relevant invited session being offered is ***Using the ACT and SAT for Accountability Under ESSA: Appropriate or Inappropriate Use*** (Sunday, 5pm).

The 2019 Annual Meeting program features 22 training sessions, 50 coordinated sessions, 7 Electronic Board sessions (2 of which are Graduate Student sessions), and over 50 individual paper sessions. Another conference highlight will certainly be the featured session for NCME's Committee on Diversity in Testing, ***Equity-Centered Design in Assessment*** (Monday, 10:25am). And we have a session devoted to recognizing the particular contributions of women to the measurement field, ***Women in Measurement: Their Unique Contributions*** (Saturday, 8am). In addition, the 2018 NCME Career Award winner, Brian Clauser, will present his research in a special session on Monday at 12:20pm, and the 2019 NCME Career Award winner, Shelby Haberman, will present his research in a special session on Sunday at 3:20pm. The 2019 NCME Award Winners will present their award-winning research on Saturday at 8am. There will also be a session focused on a new NCME award, ***Excellence in Public Communications***, that you will want to check out on Saturday at 4:10pm. If that's not enough, there's yoga on Saturday morning at 6:30am, receptions every night, the popular NCME Breakfast and Presidential Address from NCME President Rebecca Zwick on Sunday morning, and the NCME Fitness Run/Walk first thing on Monday.

We must acknowledge the incredible and talented cadre of NCME members and colleagues who generously volunteered their time and expertise this year to ensure that we will have a high-quality program in Toronto. So many of you reviewed proposals and provided critically helpful feedback; and so many of you volunteered and are serving as discussants for the program. We can't thank you enough! We are also grateful to Rebecca Zwick, our NCME President; Jim Roberts, the Training & Professional Development Committee Chair, and his committee; Kevin Krost, the Graduate Student Committee Chair, and his committee; and Nikole Gregg for their work on the program this year.

We are so excited the conference is here! Please enjoy!

Krista Mattern and Emily Shaw
2019 NCME Annual Meeting Co-Chairs

Table of Contents

NCME Officers

President	**Rebecca Zwick** *Educational Testing Service*
President-Elect	**Stephen G. Sireci** *University of Massachusetts Amherst*
Past President	**Randy Bennett** *Educational Testing Service*

NCME Directors

Derek Briggs
University of Colorado

Debbie Durrence
Gwinnett County Public Schools

Andrew Ho
Harvard University

Rose McCallin
Colorado Department of Regulatory Agencies

Ye Tong
Pearson

Walter Way
The College Board

Editors

Journal of Educational Measurement	**George Engelhard** *University of Georgia*
	Jonathan Templin *University of Iowa*
	Sandip Sinharay (incoming) *Educational Testing Service*
Educational Measurement: Issues and Practice	**Deborah Harris** *University of Iowa*
ITEMS Editor	**Andre Rupp** *Educational Testing Service*
NCME Book Series Editor	**Brian Clauser** *National Board of Medical Examiners*
NCME Newsletter Editor	**Megan Welsh** *University of California – Davis*
NCME Website Editors	**Matthew Gaertner** *WestEd*
	Brian Leventhal *James Madison University*

2019 Annual Meeting Chairs

Annual Meeting Program Chairs	**Emily Shaw** *The College Board* **Krista Mattern** *ACT, Inc.*
Graduate Student Issues Committee Chair	**Kevin Krost** *Virginia Polytechnic Institute*
Training and Professional Development Committee Chair	**James Roberts** *Georgia Institute of Technology*
Fitness Run/Walk Directors	**Jill R. van den Heuvel, Ph.D.** *Alpine Testing Solutions* **Katherine Furgol Castellano, Ph.D.** *Educational Testing Service* **Brian F. French, Ph.D.** *Washington State University*

NCME Information Desk

The NCME Information desk is located in the Concert Hall Foyer at the Fairmont Royal York.

It will be open at the following times:

Thursday, April 4 . . . . . 7:30am – 4:30pm

Friday, April 5 . . . . . 7:30am– 4:30pm

Saturday, April 6 . . . . . 8:00am – 4:30pm

Sunday, April 7. . . . . 10:00am – 4:30pm

Monday, April 8 . . . . . 8:00am – 1:00pm

Proposal Reviewers

Terry Ackerman
Bercem Akbayin
Usama Ali
Jeff Allen
Cristina Anguiano-Carrasco
Alvaro Arce
Meirav Arieli-Attali
Ben Babcock
Erin Banjanovic
Patricia Baron
Luz Bay
Jonathan Beard
Kirk Becker
Beata Beigman Klebanov
Isaac Bejar
Aarti Bellara
Dmitry Belov
Michelle Boyer
Nurliyana Bukhari
Heather Buzick
Wayne Camara
Kevin Cappaert
Michael Chajewski
Jyun-Hong Chen
Yi-Hsin Chen
Edison Choe
Youn-Jeng Choi
Man-Wai Chu
Gregory Cizek
Amy Clark
Kimberly Colvin
Stephen Cubbellotti
Zhongmin Cui
Shenghai Dai
Laurie Davis
R. De Ayala
Robert Dedrick
Christine DeMars
John Denbleyker
Ben Domingue
John Donoghue
Bryan Drost
Carol Eckerly
Howard Everson
Maureen Ewing
Fen Fan
Meichu Fan
Yu Fang
Tia Fechter
Leah Feuerstahler
Anthony Fina
Holmes Finch
Steven Fitzpatrick
Kelly Foelber
Jean-Paul Fox
John Fremer
Brian French
Hirotaka Fukuhara
Matthew Gaertner
Jennifer Galindo
Xiaohong Gao
Tracy Gardner
Joshua Goodman
Irina Grabovsky
Edith Graf
Raman Grover
Lixiong Gu
Kyung (Chris) T. Han
Qiwei Britt He
Yong He
Dianne Henderson
Amy Hendrickson
Dolores Hidalgo
TsungHan Ho
Steven Holtzman
Likun Hou
Chia-Ling Hsu
Anne Corinne Huggins-Manley
Charles Hunter
Sukkeun Im
William Insko
Paul Jewsbury
Yue Jia
Zhehan Jiang
Kwanghee Jung
Pamela Kaliski
Priya Kannan
Shu-Chuan Kao
Tzur Karelitz
Leslie Keng
Eunhee Keum
Minsung Kim
Se-Kang Kim
Seock-Ho Kim
Young Yee Kim
Tim Konold
Jason Kopp
Patrick Kyllonen
Emily Lai
Hollis Lai
Joni Lakin
Erika Landl
Quinn Lathrop
Chansoon (Danielle) Lee
Brian Leventhal
Daniel Lewis
Jie Li
Shuhong Li
Tongyun Li
Yuan-Ling Liaw
Chunyan Liu
Jinghua Liu
Junhui Liu
Yuming Liu
Samuel Livingston
John Lockwood
Susan Lottridge
Ru Lu
Ying Lu
Richard Luecht
Xiao Luo
Yong Luo
Wenchao Ma
Katerina Marcoulides
Jessica Marini
Scott Marion
Kimberly Marsh
Jose Felipe Martinez
Martha McCall
Rose McCallin
Catherine McClellan
Bradley McMillen
Maria Medina-Diaz
Rob Meijer
Gerald Melican
Yu Meng
Yeow Meng Thum
Stefan Merchant
Rochelle Michel
M. Miller
Scott Monroe
Melinda Montgomery
Joann Moore
Kristin Morrison
Eric Moyer
James Olsen
Insu Paek
Seohong Pak
Tianshu Pan
Richard Patz
John Poggio
Cornelis Potgieter
Sonya Powers
Jiahe Qian
Justine Radunzel
Heather Rickels
Michael Rodriguez
Jonathan Rubright
André Rupp
Michael Russell
Leslie Rutkowski
Edgar Sanchez
Edynn Sato
Amy Schmidt
Matthew Schultz
Bernard Schuster
Carl Setzer
S. Kanageswari Shanmugam
Can Shao
Benjamin Shear
Mark Shermis
Bruce Shotts
Sandip Sinharay
William Skorupski
Whitney Smiley
Jessalyn Smith
Hao Song
Dorota Staniewska

Proposal Reviewers (continued)

Jeffrey Steedle
Dubravka Svetina
Matthew Swain
Nadine Talbot
Wei Tao
Michael Tappler
Catherine Taylor
William Thompson
Ye Tong
Anne Traynor
Ahmet Turhan
Jon Twing
Jill van den Heuvel
Michael Walker
Aijun Wang
Lin Wang
Min Wang
Xiaolin Wang
Ze Wang
Walter Way
Jonathan Weeks
C. Whittington
Andrew Wiley
Scott Wood
Saskia Wools
Meng Wu
Yi-Fang Wu
E. Wylie
Adam Wyse
Nuo Xi
Jing-Ru Xu
Duanli Yan
Ping Yin
Hanwook Yoo
Xiaofeng Yu
Diego Zapata-Rivera
April Zenisky
Caiyan Zhang
Mo Zhang
Ou Zhang
Yu Zhang
Jishen Zhao
Xiaying Zheng
Xiaoliang Zhou
Mengxiao Zhu
Rongchun Zhu
Bruno Zumbo

Graduate Student Abstract Reviewers

Ella Banda
Yu Bao
Tanesia Beverly
Ummugul Bezirhan
Yanhong Bian
Sandra Botha
Alex Brodersen
Ian Campbell
Delwin Carter
Rajendra Chattergoon
Chia-Wen Chen
Yi-Chen Chiang
Lilian Chimuma
Dakota Cintron
Brittany Crawford
Sien Deng
Victoria Driver
Zachary Feldberg
Yanan Feng
Yan Fu
Nikole Gregg
Yage Guo
Gulsah Gurkan
Heather Handy
Minami Hattori
Maxwell Hong
Xuejun Ji
Shumin Jing
David Johnson
Unhee Ju
Hyun Joo Jung
Youngsoon Kang
Hacer Karaemse
Daniel Katz
Seohyun Kim
Sohee Kim
Kevin Krost
Minhyeong Lee
Anqi Li
Hwanggyu Lim
Ye Lin
Huan Liu
Mingjia Ma
Kaiwen Man
Kyle Nickodem
Susan Niessen
Luping Niu
Francis O'Donnell
Thai Ong
Soyoung Park
Yooyoung Park
Duy Pham
Yuxi Qiu
Daniella Rebouças
Ray Reichenberg
Aileen Reid
Jennifer Reimers
Tyler Sandersfeld
Kun Su
Victoria Tanaka
Chen Tian
Tong Wu
Qing Xie
Menglin Xu
Jiahui Zhang
Jiaqi Zhang

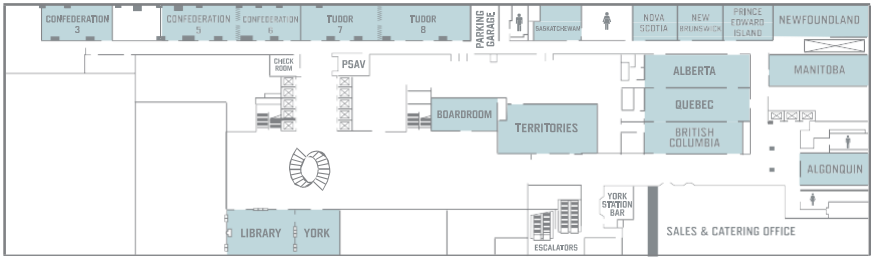
Future Annual Meetings

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

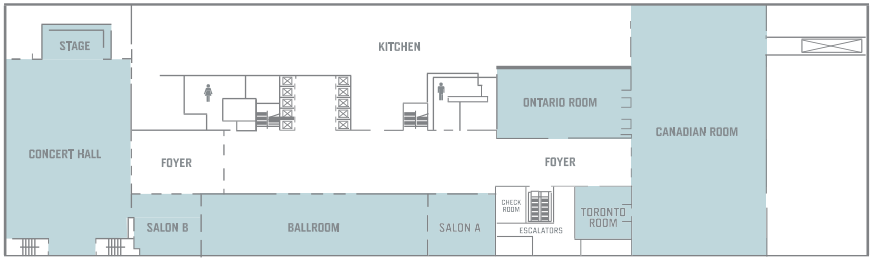
Floor Plans

Fairmont
ROYAL YORK

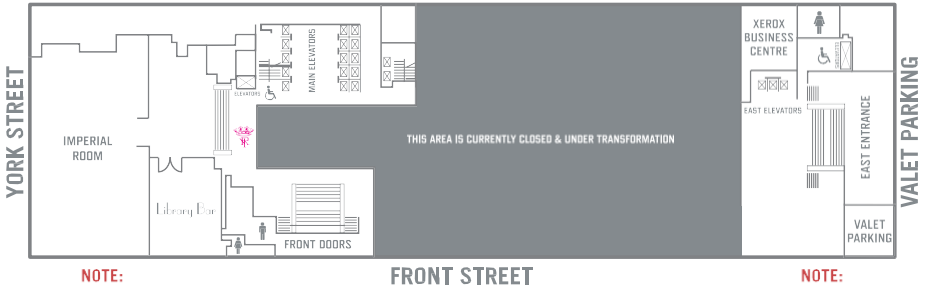
MAIN MEZZANINE



CONVENTION FLOOR



LOBBY LEVEL (MAIN FLOOR)



NOTE:
For access to Library Bar, Imperial Room, and 19th Floor use Main Elevators.

FRONT STREET

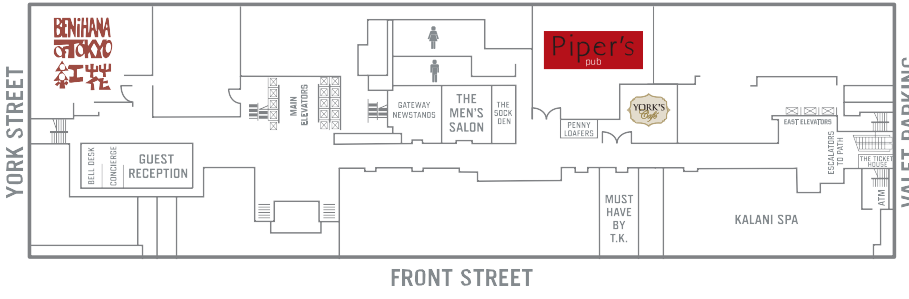
NOTE:
For access to Business Centre and Valet Parking use East Elevators.

Library Bar
Open seven days a week

À La Carte Breakfast - 7 a.m. to 10:15 a.m.
Lunch - 11:30 a.m. to 5 p.m. ▪ Monday to Friday
Afternoon Tea - Seatings from 12 p.m. to 2:45 p.m. ▪ Saturday & Sunday
Dinner - 5 p.m. to 1 a.m.

Floor Plans

AVENUES LEVEL

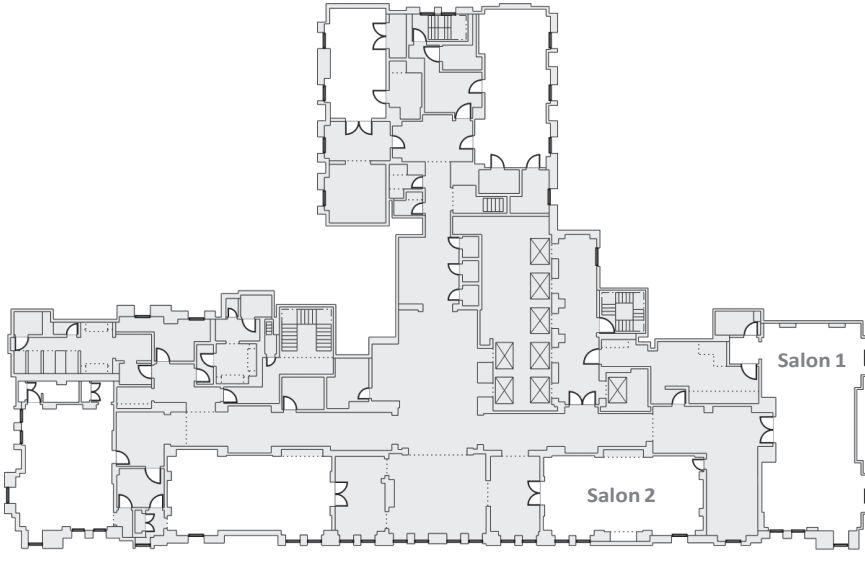


Piper's Pub
Open seven days a week
Buffet Breakfast - 7 a.m. to 11 a.m.
Lunch - 12 p.m. to 4 p.m. ▪ Saturday & Sunday
Dinner - 4 p.m. to 12 a.m.

York's Café
Open seven days a week
Light Breakfast - 6:30 a.m. to noon

Benihana Japanese Steakhouse and Lounge
Dinner - 5:30 p.m. to 10 p.m. ▪ Tuesday - Saturday

19TH FLOOR



Pre-Conference Training Sessions

The 2019 Pre-Conference Training Sessions will be held at the Fairmont Royal York on Thursday and Friday, April 4th & 5th. All full-day sessions will be held from 8:00am to 5:00pm. All half-day morning sessions will be held from 8:00am to 12:00pm. All half-day afternoon sessions will run from 1:00pm to 5:00pm. Onsite registration for the Pre-Conference Training Sessions will be available at the NCME Information desk located in the Concert Hall Foyer of the Fairmont Royal York.

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

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

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

Thursday, April 4, 2019
8:00am – 5:00pm, Alberta, Training Session

LNIRT: Joint Modeling of Accuracy and Process Data
Instructor: Jean-Paul Fox, University of Twente
Instructor: Konrad Klotzke, University of Twente

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

Thursday, April 4, 2019
8:00am – 12:00pm, Algonquin, Training Session

Introduction to R Software and Applications

Instructor: Randall E. Schumacker, The University of Alabama

The R software will be downloaded and installed by each participant. Participants will learn how to navigate the R software menu. A review of R packages and their accessibility from the pull-down menu will be covered. A review of some R functions and their arguments will be described for a few R packages. Next, participants will be shown how to write and save a Script file, which will contain specific package(s) and function(s). The training session will also cover one or more useful websites for end-users, knowledge about the UseR! Conference, and the R journal. Finally, sample Script programs with basic functions, simulations, and graphing will be demonstrated. Participants will view Instructors laptop materials on an overhead projector screen. Participants will use their laptops to follow along with the instruction topics, install, access, and save materials during the presentation. Learning objectives: Participants will be able to download and install R, navigate and use R software packages, create R script programs, access helpful R websites, knowledge of R users conference and journal, and run sample R script programs with basic functions, simulation, and graphing. Audience: Academic faculty, practitioners, psychometricians Laptop/ Software use: R software on Windows PC, Apple, or Linux laptops

Thursday, April 4, 2019
8:00am – 5:00pm, British Columbia, Training Session

Measuring Social, Emotional, and Self-Management Skills for Schools and the Workplace

Instructor: Patrick Charles Kyllonen, Educational Testing Service
Instructor: Jiyun Zu, ETS
Instructor: Jonas Bertling, Educational Testing Service

This workshop provides training, discussion, and hands-on experience in developing methods for assessing, scoring, and reporting on social-emotional and self-management skills, for K-12, higher education, and the workplace. Workshop focuses on (a) reviewing the most important skills based on current research; (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, IRT, and generalizability theory; 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., PISA, NAEP, SuccessNavigator, FACETS). There will be hands-on demonstrations using R for scoring anchoring vignettes and SJTs. 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).

Thursday, April 4, 2019
8:00am – 5:00pm, Quebec, Training Session

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

Instructor: Jimmy de la Torre, The University of Hong Kong
Instructor: Wenchao Ma, The University of Alabama – Tuscaloosa

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 (IRT) 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.

Thursday, April 4, 2019
8:00am – 12:00pm, Salon A, Training Session

NAEP Response Process Data

Instructor: Emmanuel Sikali, U.S. Department of Education
Instructor: Xiaying Zheng, American Institutes for Research
Instructor: Fusun Sahin, American Institutes for Research
Instructor: Ruhan Circi, American Institutes for Research

The National Assessment of Educational Progress (NAEP) has begun to transition to digitally-based assessments (DBA), starting with Grades 4 and 8 Mathematics and Reading in 2017. Availability of student response process data with the introduction of DBA creates numerous possibilities for psychometricians and researchers interested in examining the detailed logs of students' interactions with items and the assessment interface that can offer insight in response processes. In this session, participants will be guided on how to convert simulated NAEP response process data in raw format (XML files) into the R open source software environment step-by-step so that they can create a more accessible data format and extract/create new variables such as item response times. Participants will learn how to conduct analysis with response process data for their various research questions, through instructors' demonstrations of data analyses and visualization. Additionally, exemplary research projects using process data commissioned by NCES and conducted by AIR will be presented to participants. They range from using NAEP process data to explore students' aberrant behaviors, to the use of response process data to inform test development. Intended participants are researchers, including graduate students, education practitioners, and policy analysts, who are interested in NAEP process data analysis.

Thursday, April 4, 2019
8:00am – 5:00pm, Salon B, Training Session

Bayesian Networks in Educational Assessment (Book by Springer)
Instructor: Duanli Yan, ETS
Instructor: Russell Almond, Florida State University
Instructor: Roy Levy, Arizona State University
Instructor: Diego Zapata-Rivera, Educational Testing Service

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

Thursday, April 4, 2019
1:00 – 5:00pm, Algonquin, Training Session

Optimal Test Design Approach to Fixed and Adaptive Test Construction Using R

Instructor: Seung W. Choi, The University of Texas - Austin

In recent years, fixed 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. However, in conventional CAT items are selected through computer algorithms to meet statistical criteria along with other content-based and practical requirements, whereas fi ed forms are often created by test constructors using iterative review processes and more holistic approaches. Founded on the optimal test design framework, the shadow-test approach 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 fi ed and adaptive test construction. Practical examples will be provided along with an R package for creating and evaluating various fixed and adaptive test formats.

Thursday, April 4, 2019
1:00 – 5:00pm, Manitoba, Training Session

Using R Markdown to Automatically Generate Technical, Research, and Score Reports

Instructor: Andrew Jones, American Board of Surgery
Instructor: Carl Setzer, AICPA
Instructor: Jason P. Kopp, American Board of Surgery

Organizations and researchers often create reports and manuscripts by manually copying information from statistical software output into a document. Such processes are inefficient and are susceptible to error. The purpose of this session is to demonstrate R's capability to generate automated reports for technical documentation, research manuscripts, item analysis reports, and examinee performance reports. We will utilize the statistical software R, through R-Studio, to demonstrate report generation using R Markdown with KnitR and LaTeX. The workshop will be hands-on and participants will run their own reports.

Thursday, April 4, 2019
1:00 – 5:00pm, Salon A, Training Session

Analyzing Features of Assessment Items: An Introduction

Instructor: Jenny C. Kao, UCLA/CRESST
Instructor: Elizabeth Redman, UCLA/CRESST
Instructor: Kilchan Choi, UCLA/CRESST

Participants will be introduced to feature analysis, a process in which assessment items are qualitatively rated with a set of attributes, followed by subsequent quantitative analysis 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 that are suitable for instruction, and providing a method for comparability across assessment items. Participants will have the opportunity to engage in and practice rating test items—both traditional multiple-choice math and English language arts as well as computerized, game-based assessments. Participants will learn about specific, qualitative features of test items that will inform both future test development, interpretation of test scores, and any potential for item bias. This session will be of interest to researchers and test developers who would like to broaden their understanding of how test item features beyond content/domain interact with student performance, as well as to practitioners and policymakers interested in making inferences about test scores beyond content or domain.

Friday, April 5, 2019
8:00am – 5:00pm, Alberta, Training Session

Generalizability Theory and Applications
Instructor: Robert L. Brennan, University of Iowa
Instructor: Won-Chan Lee, University of Iowa

Generalizability theory liberalizes and extends classical test theory. In particular, generalizability theory enables an investigator to disentangle multiple sources of error through the application of analysis of variance procedures to assess the dependability of measurements. The primary goals of this training session are to enable participants to understand the basic principles of generalizability theory, to conduct relatively straightforward generalizability analyses, and to interpret and use the results of such analyses. Mathematical and statistical foundations will be treated only minimally. Major emphasis will be placed upon quickly enabling participants to conduct and interpret relatively straightforward generalizability analyses, then more complicated ones. Examples will include various types of performance assessments. Prerequisites include knowledge equivalent to one course in educational measurement and familiarity with ANOVA at an introductory level."Generalizability Theory," a book written by the senior director, will be distributed to participants. Computer programs for performing generalizability analyses will be discussed and illustrated. (Participants need not bring laptops.)

Friday, April 5, 2019
8:00am – 12:00pm, Algonquin, Training Session

A Visual Introduction to Computerized Adaptive Testing

Instructor: Yuehmei Chien, NWEA
Instructor: Ching-Wei D Shin, Pearson

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

Friday, April 5, 2019
8:00am – 5:00pm, British Columbia, Training Session

Learning More From Test Data: New Tools for Test Scoring

Instructor: James Ramsay, McGill University
Instructor: Juan Li, McGill University
Instructor: Marie Wiberg, Umeå University

The aim of scoring a test is to give as best estimate of an examinee's ability as possible. The goals of this training session are for the attendees to be able to understand and implement optimal test scoring, and to interpret the results of optimal scoring in a reasonable way. In this training session, we will demonstrate and guide the attendees to use the web-based software TestGardener to implement optimal test scoring on real educational test data. Most of the outputs of this software are in graphical form, and the software is used interactively. The main part of the training session is devoted to practical exercises in how to analyze test data. Optimal scoring will also be compared with the traditional sum scoring, and recent developments in test scoring will be discussed. Expected audience include researchers, graduate students and practitioners. An introductory statistical background is recommended but not required. Please note, programming knowledge is not required.

Friday, April 5, 2019
8:00am – 12:00pm, Manitoba, Training Session

Tips and Tricks to Effectively Communicate Results: Best Practices in Data Visualization

Instructor: Nikole Gregg, James Madison University
Instructor: Brian Leventhal, James Madison University

Communication to a general audience, including educators, stakeholders, and students, is a necessary skill for educational measurement researchers and practitioners. Data visualization and graphical excellence are necessary to communicate interpretable data effi tly and truthfully. Measurement and statistical software commonly produce graphics by default, however, they are typically not suitable for publication or presentation to the public. Use of SAS has several advantages: 1) it is commonly used across disciplines; 2) it provides a robust programming; 3) it allows customization options to suit multiple outlets (i.e. journals, presentations, etc.); and 4) it facilitates input of data/results from specialized measurement software. This training session illustrates best practices of data visualization and how the SAS template language can be used to produce presentation and journal quality graphics. Through demonstration, application, and active learning, attendees will understand the basic components of sound data visualization, identify misleading and inaccurate graphics, and have a base knowledge of the SAS template language.

Friday, April 5, 2019
8:00am – 12:00pm, Quebec, Training Session

Computerized Multistage Adaptive Testing: Theory and Applications (Book by Chapman and Hall)
Instructor: Duanli Yan, ETS
Instructor: Alina A. Von Davier, ACT, Inc.
Instructor: Kyung (Chris) T. Han, The Graduate Management Admission Council

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

Friday, April 5, 2019
8:00am – 5:00pm, Salon A, Training Session

Exploring, Visualizing, and Modeling Big Data With R

Instructor: Okan Bulut, University of Alberta
Instructor: Christopher David Desjardins, University of Minnesota

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, support vector machines, and k-nearest neighbors), as well as hands-on demonstrations applying these methods in R. Examples will be drawn from various assessments (e.g., PISA, TIMSS, and NAEP). 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.

Friday, April 5, 2019
8:00am – 12:00pm, Salon B, Training Session

Nonparametric Cognitive Diagnosis and Computer Adaptive Testing for Small Samples

Instructor: Chia-Yi Chiu, Rutgers
Instructor: Hans Friedrich Köhn, University of Illinois at Urbana-Champaign

The training sessions concern methods of cognitive diagnosis that are tailored to the use in small-scale educational settings like the classroom, where the number of examinees is simply too small so that parameter-based estimation methods (e.g., marginal maximum likelihood estimation relying on the Expectation Maximization algorithm or Markov chain Monte Carlo techniques) fail in analyzing the data. Nonparametric methods are presented as an alternative to parameter-based estimation for analyzing assessment data from small-scale educational settings within the cognitive diagnosis framework. Four sessions address the following topics: (1) construction of complete Q-matrices (a complete Q-matrix is the core of any cognitively diagnostic test); (2) nonparametric methods for Q-matrix validation; (3) nonparametric classification methods for cognitive diagnosis in small educational settings; (4) nonparametric Computerized Adaptive Testing for cognitive diagnosis in small educational settings. The goal of the training sessions is to familiarize participants with recently developed nonparametric methods for cognitive diagnosis and to provide hands-on training in the R programs implementing these methods. The training sessions are of interest to anyone who wishes to use or research cognitive diagnosis in small-scale educational settings. Basic knowledge in Item Response Theory and prior exposure to R would be helpful, but are not strict requirements.

Friday, April 5, 2019
1:00 – 5:00pm, Algonquin, Training Session

Using SAS for Monte Carlo Simulation Studies in Item Response Theory

Instructor: Brian Leventhal, James Madison University
Instructor: Allison Ames, University of Arkansas

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.

Friday, April 5, 2019
1:00 – 5:00pm, Confederation 6, Training Session

Diagnostic Classification Models Part II: Advanced Applications

Instructor: Matthew James Madison, Clemson University

Diagnostic measurement is an emerging field of psychometrics that focuses on providing actionable feedback from multidimensional tests. This workshop provides a more advanced introduction to diagnostic classification models (DCMs). More specifically, it focuses on the structural component of DCMs, estimation using R, and recent advancements in longitudinal DCMs. After completing this workshop, participants will understand the statistical structure of DCMs, be able to estimate DCMs and interpret software output, and understand how longitudinal DCMs can be applied to assess change in mastery profi over time. This session is appropriate for graduate students, researchers, and practitioners at the emerging or experienced level. Participants are expected to have only a basic knowledge of statistics and psychometrics to enroll. This session presents both conceptual and technical content and also provides hands-on experience for participants to apply what they learn. Material is presented at a technical level when necessary for understanding the models and applying them responsibly. Content will mostly be delivered through lecture, and content will be reinforced using hands-on activities. Instructors will encourage audience participation through questions and allow time for discussions among participants and the instructors.

Friday, April 5, 2019
1:00 – 5:00pm, Manitoba, Training Session

Vertical Scaling Methodologies, Applications, and Research

Instructor: YeTong, Pearson
Instructor: Michael J. Kolen, The University of Iowa

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

Friday, April 5, 2019
1:00 – 5:00pm, Quebec, Training Session

Tools and Strategies for the Design and Evaluation of Score Reports

Instructor: Diego Zapata-Rivera, Educational Testing Service
Instructor: Priya Kannan, Educational Testing Service
Instructor: Sharon Cadman Slater, ETS
Instructor: April L. Zenisky, University of Massachusetts - Amherst
Instructor: Gavin T. Brown, The University of Auckland

Score reports are often the primary means by which score users receive information about tests. Score report users often have different levels of familiarity with and understanding of not only the assessment but also the psychometrics behind the scores reported. Therefore, 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. In alignment with this year's conference theme of "Communicating with the Public about Educational Measurement", this training session offers 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 a combination of an interactive lecture format and hands-on practical experience, and equip the attendees with various hand-outs of 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 session. This session is based on the recent NCME Book on "Score Reporting Research and Applications" – a copy of the book will be included as part of session registration.

Friday, April 5, 2019
1:00 – 5:00pm, Salon B, Training Session

An Introduction to the Use of Telemetry Data in Video Game Analyses
Instructor: Gregory Chung, CRESST
Instructor: Tianying Feng, University of California - Los Angeles
Instructor: Charlie Parks, University of California - Los Angeles
Instructor: Elizabeth Redman, UCLA/CRESST
Instructor: Jeremy Roberts, PBS KIDS Digital
Instructor: Katerina Schenke, University of California - Los Angeles

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 specifi 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.

Friday, April 5, 2019
1:00 – 5:00pm, Territories, Training Session

Software Packages for Item Response Theory–Based Test Simulation: WinGen3, SimulCAT, MSTGen, and IRTEQ

Instructor: Hanwook (Henry) Yoo, Educational Testing Service
Instructor: Kyung (Chris) T. Han, The Graduate Management Admission Council
Instructor: Hyeonjoo J. Oh, ETS

This training session introduces four item response theory (IRT)-based simulation computer programs (1) WinGen3 for generating IRT parameters and item responses, (2) SimulCAT for simulating computer adaptive testing administrations, (3) MSTGen for simulating multistage testing administrations, and (4) IRTEQ for implementing IRT equating. These software tools support various IRT models and comprehensive features with intuitive, user-friendly interface. Out of this training session, attendees will have better understanding of the importance of IRT-based simulation as well as the practical constraints and challenges of simulation-based research. The current training delivers essential psychometric knowledge and professional simulation skills, as well as passes down the practical tips to write well-defined and impactful research questions for simulation study. The workshop is intended for junior-level practitioners and graduate students. It is recommended for participants to have some background knowledge in modern test theory (a.k.a., IRT) including differential item functioning, item parameter drift, scaling and equating, and multistage testing issues but not required. Demonstrations and hands-on practice will be conducted with proposed free software programs. Attendees should bring their own laptops and the most recent version of three programs installed (www.hantest.net). Presenters will send electronic training materials via email at least one week prior to the conference.

Saturday, April 6, 2019
8:00 – 10:00am, Alberta, Invited Speaker Session

Women in Measurement: Their Unique Contributions

Chair: Linda L. Cook, Educational Testing Service
Discussant: Mary Pitoniak, Educational Testing Service

The contributions of women have been minimized or overlooked in the histories of many professions. The field of measurement is no exception. The purpose of this session is to highlight the contributions of women from four different sectors in our field: academic institutions, test publishing organizations, professional organizations, and federal and state organizations. This session will explore each one of these sectors and describe the unique ways in which women have helped the field to progress. Both seminal contributions made in the middle and latter parts of the 20th century and those made more recently will be reviewed. Attention will also be paid to the potential for women's future contributions in these four sectors.

Federal and State Organizations
Peggy G. Carr, National Center for Education Statistics/IES, U.S. Department of Education

Academic Institutions
Kadriye Ercikan, Educational Testing Service, Princeton, NJ 08541; Han-Hui Por, Educational Testing Service

Professional Organizations
Joan Herman, University of California Los Angeles/CRESST

Test Publishing Organizations
Ida M. Lawrence, ETS; Edward Shea, ETS

Saturday, April 6, 2019
8:00 – 10:00am, Algonquin, Paper Session

Advances in Cognitive Diagnostic Modeling

Discussant: Howard T. Everson, SRI International & City University of New York

Estimation of Partially Defined Q-Matrix

Qianru Liang, The University of Hong Kong; Jimmy de la Torre, The University of Hong Kong

We propose an estimation method for partially defined Q-matrix based on the likelihood of all possible q-vectors for each item. The algorithm chooses the q-vector with the smallest information criterion. Results indicate that it performs well in terms of element-wise recovery rate for large sample sizes and high item quality.

Multiple-Strategy Cognitive Diagnosis Models for Dichotomous Response

Wenchao Ma, The University of Alabama - Tuscaloosa; Wenjing Guo, The University of Alabama

This study develops a generalized multiple-strategy cognitive diagnosis model for dichotomous response. The model provides a unified framework to accommodate various condensation rules and strategy selection approaches. Simulation studies showed that the parameters of the proposed model can be adequately recovered and that the proposed model was relatively robust.

Multilevel Analysis Incorporating Multiple Covariates for Independent and Higher Order Cognitive Diagnosis Models

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

This study proposed a generalized approach to incorporate multiple covariates for cognitive diagnostic models (CDMs) with multilevel data structures. Real-world data analysis using large-scale multilevel data were used to demonstrate the applications of the approach. Simulation studies were conducted to examine the consistency of parameter recovery and estimation.

Improving Classification Accuracy in High Dimensional Data: A Three-Step Approach

Yan Sun, Rutgers University - New Brunswick/Piscataway; Jimmy de la Torre, The University of Hong Kong

In this study, covariates are incorporated to improve classifi tion accuracy in large dimensional cognitive diagnostic tests. The performance of the proposed approach was examined in a simulation study. Results showed that the proposed approach could increase information obtained from CDM and improve the classifi tion accuracy when tests are not informative.

On the Equivalence of Unidimensional Item Response Theory and Cognitive Diagnosis Models

Jimmy de la Torre, The University of Hong Kong; Kevin Carl Pena Santos, University of the Philippines

Educational assessments developed using unidimensional item response theory framework have been retrofitted with cognitive diagnosis models to generate more diagnostic feedback. However, it remains unclear how the two disparate psychometric frameworks can be simultaneously used to analyze the same assessment data. This paper proposes a unifying framework for such applications.

Saturday, April 6, 2019
8:00 – 10:00am, Ballroom, Coordinated Session

Applications of Multilevel Item Response Theory Models for Collecting Validity Evidence in Educational Assessments

Chair: Jan Hochweber, University of Teacher Education St.Gallen
Discussant: Guillermo Solano-Flores, Stanford University

Educational assessments are widely used within schools, educational research and evidence-based policy making. For instance, tests are commonly applied tools for assessing individual students' competencies or achievement. On the aggregate level, test scores have become a major criterion for determining the effectiveness of teaching and schooling. However, claims on student learning or teaching effectiveness based on tests require establishing links between the empirical data and the inferential target as validity evidence. In recent years, multilevel item response (MLIRT) models have become a powerful tool in educational research for collecting such validity evidence. MLIRT models link item responses to latent variables while allowing accounting for hierarchical data structures with, for example, students nested in classes, courses or schools. The proposed symposium comprises four presentations related to innovative applications of MLIRT models to collect validity evidence in educational assessments addressing (a) cluster-level dimensionality of measures of students'learning, (b) a longitudinal multilevel extension of the linear logistic test model (LLTM) to identify item properties related to instructional sensitivity. (c) using MLIRT models as a validation strategy for expert judgements on items' instructional sensitivity, and d) using MLIRT to deal with measurement error when investigating the relationship of variables at different levels.

A Longitudinal Multilevel Item Response Theory Model to Evaluate Multidimensionality of Change or Change in Dimensionality at the Cluster Level

Alexander Naumann, DIPF | Leibniz Institute for Research and Information in Education; Johannes Hartig, German Institute for International Educational Research

A Longitudinal Multilevel Extension of the Linear Logistic Test Model to Predict the Instructional Sensitivity of Test Items

Jan Hochweber, University of Teacher Education St.Gallen; Alexander Naumann, DIPF | Leibniz Institute for Research and Information in Education; Iris Kleinbub, University of Education Ludwigsburg; Johannes Hartig, German Institute for International Educational Research; Stephanie Musow, University of Teacher Education St. Gallen

Multilevel Item Response Theory as a Validation Strategy for Expert Judgments on Instructional Sensitivity

Stephanie Musow, University of Teacher Education St. Gallen; Alexander Naumann, DIPF | Leibniz Institute for Research and Information in Education; Jan Hochweber, University of Teacher Education St.Gallen; Johannes Hartig, German Institute for International Educational Research

Testing the Generalization to the Domain Inference: The Use of Contextualized Clusters of Items

Maria Araceli Ruiz-Primo, Stanford University; Min Li, University of Washington; Jim Minstrell, FACET Innovations; Xiaoming Zhai, Stanford University; Dongsheng Dong, University of Washington - Seattle; Klint Kanopka, Stanford University; Philip Hernandez, Stanford University

Saturday, April 6, 2019
8:00 – 10:00am, British Columbia, Coordinated Session

Technology-Enhanced Items: Lessons Learned and Future Directions

Chair: Yue Jia, Educational Testing Service
Discussant: Michael C. Rodriguez, University of Minnesota

As assessments move from traditional paper-pencil administration to computer-based administration, many testing programs are incorporating technology enhanced items (TEIs) into assessments with the goals of measuring higher-order thinking, offering insight into problem-solving, and representing authentic real-world tasks. This session explores diverse applications of technology enhanced items and describes lessons learned from their administration. Additionally, this session introduces methodologies intended to model aspects of TEIs which are not considered in traditional latent variable models. The session includes commentary and discussion from an expert in item development.

Technology-Enhanced Items: Signal or Noise?
Wayne J. Camara, ACT, Inc.; Wei Tao, ACT, Inc.

Technology-Enhanced Items and Model-Data Misfit
Carol Eckerly, Educational Testing Service; Yue Jia, Educational Testing Service

Worth the Squeeze? An Investigation into the Psychometric Performance of Innovative Item Types
Amanda A. Wolkowitz, Alpine Testing Solutions; Brett Patrick Foley, Alpine Testing Solutions

Using Asymmetric Item Response Theory Models to Accommodate Item Complexity
Daniel M. Bolt, University of Wisconsin - Madison; Sora Lee, The University of Wisconsin - Madison; James A. Wollack, University of Wisconsin - Madison; John Sowles, Ericsson

Event History Analysis of Process Data From PSTRE
Zhiliang Ying, Columbia University

Saturday, April 6, 2019
8:00 – 10:00am, Manitoba, Paper Session

Practical Applications of Validity Research

Discussant: Paul Westrick, The College Board

Examining Construct Shift in Reading Development

John P. Sabatini, ETS; Jonathan P. Weeks, Educational Testing Service; Tenaha P. O'Reilly, ETS; Zuowei Wang, Educational Testing Service

We examined construct shift between low, average, and high skilled readers on a battery of reading measures. Based on reading theory, we predicted and found that a multiple factor model of reading ability fit better for low skilled readers, whereas a single factor model fit better for high skilled readers.

Optimal Methods for Disattenuating Correlation Coefficients Under Realistic Measurement Conditions

Carrie Morris, ACT; Walter Peter Vispoel, University of Iowa

Commonly used methods for disattenuating correlations fail to account for between occasion error and correlations involving such errors. We compared multiple CFA and formula-based approaches for estimating disattenuated correlations using two occasion simulated data. The findings highlight the importance of accounting for all major sources of measurement error when disattenuating.

Identifying Core Competencies of 21st Century Learning Outcomes: A Web Scraping Approach

Joseph Rios, University of Minnesota; Guangming Ling, ETS; Robert Pugh, ETS; Dovid Becker, Educational Testing Service

This study identified core competencies of 21st century learning outcomes by conducting content analyses of 142,000 job advertisements in the U.S. economy. Results demonstrated that the most requested competencies were: oral and written communication, collaboration, and problem solving. Differences were observed by degree-level and academic field.

Evaluating Content-Related Validity Evidence Using Text Modeling

Daniel John Anderson, University of Oregon; Brock Rowley, University of Oregon; Sondra Stegenga, University of Oregon

Topic modeling is applied with science content standards to evaluate semantic clustering. The probability that each item from a statewide assessment belongs to each cluster/topic is then estimated as a source of content-related validity evidence. We also show how visualizations can map the content coverage of the test.

Thinking About Claims for Assessments of the Next Generation Science Standards

Mary Norris, Virginia Polytechnic Institute and State University; Brian Gong, Center for Assessment; Mary Norris, Virginia Polytechnic Institute and State University

Our work provides guidance for creating claims about student performance on the Next Generation Science Standards (NGSS). We present seven aspects of quality for describing performance, analyze types of claims that can be supported, and illustrate the application of novice/expert science research literature to create sample claims about the NGSS.

Saturday, April 6, 2019
8:00 – 10:00am, Quebec, Paper Session

Advances in the Evaluation of Item Response Theory Models

Discussant: Brian Leventhal, James Madison University

Longitudinal Randomized Controlled Trials With Item Response Data

Marian Strazzeri, University of Maryland College Park; Ji Seung Yang, University of Maryland

Maximum likelihood estimation of multiple-group latent growth models with categorical indicators becomes challenging as the number of latent variables increases. A Monte Carlo simulation study, motivated by empirical data, is conducted to evaluate the performance and practicality of fi e approaches to analyzing such data, including single- and multi-staged methods.

Power Divergence Family of Tests for Person Parameter in Item Response Theory Models

Xiang Liu, Teachers College, Columbia University; James Yang, Teachers College, Columbia University; Hui Soo Chae, Teachers College, Columbia University; Gary J. Natriello, Teachers College, Columbia University

We generalize the PD family of statistics to the IRT models. A moment matching method is introduced to choose the optimal λ within the PD family. The finite sample type I error rate, coverage rate of confidence intervals, and their lengths are evaluated via simulations. Real data examples are presented.

Exploring Psychometric Models for Process Data From Computer-Based Simulations

Yanyan Tan, University of Georgia; Matthias Von Davier, National Board of Medical Examiners; Polina Harik, National Board of Medical Examiners

This study explores psychometric models for process and choice data from computer-based case simulations in a large-scale licensure examination. Diagnostic Classification Models and Rasch Poisson Counts Model were applied to the data. Results suggest that the Rasch Poisson Counts Model extracts more reliable estimates than previously considered approaches.

Implementing Mixture Item Response Theory Models in Model Comparisons Using NUTS

Rehab Said Al Hakmani, Southern Illinois University - Carbondale; Yanyan Sheng, Southern Illinois University - Carbondale

The focus of this study is to evaluate the performance of NUTS in fitting a mixture IRT model while comparing it to the conventional IRT model using fully Bayesian fit indices. Monte Carlo simulations were carried out to compare these models under different situations and some guidelines are provided.

A Comparative Study of Item Response Theory Models for Rater Effects and Double Ratings

Yoon Ah Song, The University of Iowa; Won-Chan Lee, University of Iowa; Brandon LeBeau, The University of Iowa

This simulation study is to compare the relative performance of polytomous IRT models in dealing rater effects and double ratings. The generalized partial credit model and the hierarchical rater model were compared in terms of accuracy, test characteristic curves, and standard errors under varying combinations of factors.

Saturday, April 6, 2019
8:00 – 10:00am, Salon A, Invited Speaker Session

2019 NCME Awards Session
Chair: Denny Way, College Board

Public Communications Award
Catherine Gewertz, Education Week

Jason Millman Promising Measurement Scholar Award
Qiwei He, Educational Testing Service

Bradley Hanson Award for Contributions to Educational Measurement
Hongwen Guo, ETS

Alicia Cascallar Award
Scott Monroe, University of Massachusetts, Amherst

Annual Award
Robert J. Mislevy, Educational Testing Service

Brenda H. Loyd Outstanding Dissertation
Maria Bolsinova, ACT, Inc.

Saturday, April 6, 2019
8:00 – 10:00am, Salon B, Paper Session

Emerging Research in Multistage Testing

Discussant: Duanli Yan, ETS

Utilizing Interval Estimation and Response Time in On-the-Fly Multistage Testing

Yang Du, University of Illinois at Urbana - Champaign; Anqi Li, University of Illinois at Urbana - Champaign; Hua-Hua Chang, Purdue University

This research examines how interval estimation and response-time-based item selection methods can jointly shorten the test time while maintaining the estimation accuracy in on-the-fl multistage testing. Specifi , modules are assembled based on response time and confi e intervals of examinees' provisional ability. A comparison with computerized adaptive testing is also conducted.

Realistic Simulation for Multistage Adaptive Testing With Multiple Scales

Jennifer Reimers, University of Arkansas at Fayetteville; Sunhee Kim, College Board; Denny Way, College Board; Priyank Patel, The College Board

This simulation study investigates MST that measures multiple score scales using 3PL IRT model and a correlated multidimensional structure of multiple scales from a real testing application. The review of MST—comparison with a conventional test and exploration of reduced numbers of items in panels—suggests advantages of MST.

Handling Local Item Dependency of Testlets in Multistage Testing

Hyung Jin Kim, University of Iowa; Ah-Young Shin, American Institutes for Research

For multistage testing with testlets, three approaches can be considered to handle the local item dependency (LID) by ignoring or incorporating the presence of LID. This study investigates the effect of approaches to handling LID on the performance of the testlet-based MST in terms of measurement accuracy and decision consistency.

Ignorability of Missing Data in Multidimensional Multistage Tests

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

When modeling multidimensional multistage tests (MSTs), separate IRT models for each dimension produce unacceptable item parameter estimates. With simulations and mathematical proofs, missing data in multidimensional MSTs is shown to be non-ignorable with separate IRT models but ignorable with multidimensional IRT (MIRT).

A New Approach to Find Optimal Design of Multistage Tests

Hwanggyu Lim, University of Massachusetts; Tim Davey, ETS

The conditions under which multi-stage tests (MSTs) are applied vary case by case. An MST with optimal measurement properties in one case will not generalize to all. We propose a process for discovering an MST design with measurement properties that are optimal in some sense, given a specific test condition.

Saturday, April 6, 2019
8:00 – 10:00am, Territories, Paper Session

Pioneering Work in AIG

Discussant: Andreas H. Oranje, Educational Testing Service

Estimating Item Family Variation of Automatic Generated Items With Item Response Models

Fen Fan, National Commission on Certifi ation of Physician Assistants; Joshua T. Goodman, National Commission on Certification of Physician Assistants

Automatic item generation is one solution to meeting the content needs of high-volume testing programs. AIG produces many related items (i.e., an item family), which can complicate traditional field-test designs. The purpose of this study is to explore different statistical models for carrying out efficient field testing of item families.

Distractor Suites: A Method for Automatically Generating Answer Choices in Multiple-Choice Items

Audra Kosh, Edmentum

This presentation provides an overview of the unique challenges associated with developing answer choices for automatically-generated mathematics items and then illustrates a methodological innovation known as the distractor suite that can improve the quality of automatically-generated items while simultaneously reducing the time and effort to write and review answer choices.

Defining Crisis Papers: From Qualitative Methods to Neural Networks

Amy Burkhardt, University of Colorado - Boulder; Sherri Woolf, AIR; Christopher Ormerod, American Institutes for Research; Sue Lottridge, American Institutes of Research

We describe the challenges inherit to defining crisis papers and detail our development of an empirically supported tiered definition. We then report on how well both human raters and neural networks can distinguish between these tiers and make recommendations for improving the automatic detection of such student responses.

Automated Item Generation: Does It Pass Muster?

Cecilia Brito Alves, Medical Council of Canada; André F. De Champlain, Medical Council of Canada; Nicole Robert, Medical Council of Canada

The purpose of this study is to investigate whether one AIG process implemented in a large-scale medical licensing program yields items that meet the standards of quality expected of traditionally developed items. Quality was evaluated through classical, IRT, and generalizability (G) study analyses.

Measuring In-Context Vocabulary: Implications for Automated Item Generation

Isaac I. Bejar, ETS; Michael Flor, Educational Testing Service; Paul Deane, Educational Testing Service; Steven L. Holtzman, ETS; James Bruno, ETS

The psychometric equivalence of in-context vocabulary items and corresponding decontextualized vocabulary items was evaluated. The decontextualized items were generated by AIG while the corresponding in-context items were produced by subject matter experts. The results show that difficulty was reasonably equivalent but discrimination less so.

Saturday, April 6, 2019
10:25 – 11:55am, Alberta, Coordinated Session

Research on Test-Taking Motivation: Implications for Test Development and Educational Policy

Chair: Joseph Rios, University of Minnesota

Test-taking motivation has been shown to be a threat to the validity of score-based inferences from low-stakes assessments, which are becoming increasingly prevalent in K-16 accountability assessment contexts (e.g., PISA). The four presentations in this symposium will shed light on this important topic by focusing on the impact of ignoring low test-taking motivation on test-based educational policy decisions and examining the utility of technology-enhanced or game-based assessments in improving test-takers' motivation. Presentation 1 examines the effect of test disengagement on estimates of school contributions to student growth, achievement gaps, and summer learning loss. Presentation 2 explores data from an international assessment to evaluate the degree of differential noneffortful responding between countries and its impact on comparisons of country-level aggregated scores. Presentation 3 investigates the tenability of the long-standing assumption that technology-enhanced items are more engaging than standard text-based multiple-choice items. The last presentation employs a multi-method approach in gathering response process information to evaluate whether a game-based assessment is more motivating than a traditional computer-based assessment. The four presentations jointly advance our understanding of the impact of motivation on test scores in domestic and international samples, and test development strategies that can be used to enhance motivation.

Examining the Impact of Test Disengagement on Estimates of Educational Effectiveness
Megan Kuhfeld, NWEA; James Soland, NWEA

Is There Differential Noneffortful Responding Between Countries on an International Assessment?
Joseph Rios, University of Minnesota; Hongwen Guo, ETS

The Impact of Technology-Enhanced Items on Test-Taker Engagement
Steven L. Wise, Northwest Evaluation Association; James Soland, NWEA; Laurence Dupray, NWEA

Understanding the Motivated Test-Taking Experience to Develop More Engaging Assessments
Blair Lehman, Educational Testing Service; Tanner Jackson, Educational Testing Service

Saturday, April 6, 2019
10:25 – 11:55am, Algonquin, Coordinated Session

Advancing the Measurement Field With Data Science
Chair: Matthew Schultz, AICPA
Discussant: Carl Setzer, AICPA
Panelist: Michael C. Edwards, Arizona State University
Panelist: Peter W. Foltz, Pearson
Panelist: Sue Lottridge, American Institutes of Research
Panelist: Mark Shermis, University of Houston-Clear Lake
Panelist: Alina A. Von Davier, ACT, Inc.

The measurement field now has many decades of experience and expansion, both in terms of methods and theory. The result is a relatively entrenched and stable framework of operational test development, scoring, and scaling. Measurement, whether cognitive or non-cognitive, generally relies on either Classical Test Theory (CTT) and/or Item Response Theory. Outside of the measurement community, there are emerging fields that have impacted various research professions. Data science is a broad term for these fields. The measurement field has already seen some impact by way of automated essay scoring, for example. However, as data sciences continue to have impacts in unseen ways, the measurement community should initiate a dialog regarding how, and where, data sciences might contribute to the field, and start to develop best practices and guidelines. The purpose of this session is to discuss the potential merging of data science and measurement communities. The panel of experts will address specific questions related to the measurement community's willingness and readiness for any impacts, where these changes will be embraced, and what the measurement community can do to prepare.

Saturday, April 6, 2019
10:25 – 11:55am, Ballroom, Coordinated Session

A Tricky Balance: The Challenges and Opportunities of Balanced Systems of Assessment

Chair: Scott F. Marion, National Center for the Improvement of Educational Assessment, Inc.

Discussant: James W. Pellegrino, University of Illinois at Chicago

The seminal publication, Knowing What Students Know, crystalized the call for balanced assessment systems. Almost 20 years have passed, but still there are very few examples of well-functioning systems, particularly systems that incorporate state summative tests. Why? In spite of recent efforts to articulate principles of assessment systems, creating balanced assessment systems is really hard! This symposium builds on the work over the last 20 years to identify some high-leverage strategies that can increase the likelihood of seeing high-quality balanced assessment systems implemented in practice. The session begins with a conceptualization and defi of balanced assessment systems, with an emphasis on the factors that have likely prevented implementation. We discuss five critical factors that hinder assessment system design and implementation: (1) The politics and policy of assessment systems; (2) The commercialization, proliferation, and incoherence of assessments; (3) Interim assessments and modularity; (4) Lack of attention to learning and curriculum; and (5) Low levels of assessment literacy.

The Politics and Policy of Assessment System Design
Scott F. Marion, National Center for the Improvement of Educational Assessment, Inc.

Commercialization, Proliferation, and Incoherence of Assessments
Joseph A. Martineau, Center for Assessment

Some Emerging Design Criteria for Interim Assessments in a Balanced Assessment System
Nathan Dadey, The National Center for the Improvement of Educational Assessment, Inc.

Curriculum and Instruction in a Balanced Assessment System
Jeri Thompson, National Center for the Improvement of Educational Assessment, Inc.

Assessment Literacy: What Do Stakeholders Really Need to Know and Be Able to Do?
Carla Evans, Center for Assessment

Saturday, April 6, 2019
10:25 – 11:55am, British Columbia, Coordinated Session

Scaling Up Assessment Literacy in Teacher Preparation Programs: A Panel Discussion

Chair: Mary Yakimowski, Samford University, Alabama
Discussant: Mary Yakimowski, Samford University, Alabama
Panelist: Dorothea M. Anagnostopoulos, The University of Connecticut
Panelist: Audrey Amrein-Beardsley, Arizona State University
Panelist: Colleen Thornton MacKinnon, Independent Consultant
Panelist: James H. McMillan, Virginia Commonwealth University
Panelist: W. James Popham, University of California – Los Angeles

Stiggins (2015) and others have examined how classroom assessment literacy is taught within teacher education programs (TEPs). Now that policy-makers, such as accreditation organizations, require demonstrating use of valid and reliable assessments—thereby making demands for data literacy as part of the assessment literacy process--what is the current state and what are the areas for improvement? How can NCME assist with the latter? Has human capacity development around assessment literacy kept pace? This session brings together a diverse stakeholder panel representing those involved in TEPs, representing faculty, administration, assessment/measurement, policy, and other perspectives. This stakeholder panel will discuss the (1) assessment literacy strengths offered by TEPs, (2) obstacles and/or areas for growth needed, and (3) resources needed to help address gaps in assessment literacy expertise. In advance of the submission of this proposal, panelists were asked to submit their ideas about how NCME may further assist in this area. The panelists' aggregated preliminary list will be shared with audience participants. Then, one or more of the panelists will discuss the rationale for the recommendation. It is hoped that the audience will, in turn, share their perspectives regarding what can be offered.

Saturday, April 6, 2019
10:25 – 11:55am, Imperial Room, Electronic Board Session

Electronic Board Session 1

Model-Based Approaches to Subgroup Analysis in Automated Scoring
Scott William Wood, ACT, Inc.

The Standards for Educational and Psychological Testing encourage professionals to conduct subgroup analyses to show a lack of bias in automated scores. Methods included scoring metric analysis at the subgroup level and differential feature functioning. This presentation considers a new approach: using modified automated scoring models to identify possible bias.

Cross-Informant Associations of School Characteristics: A CT-C(M-1) Analysis of Students and Adults
Timothy R. Konold, University of Virginia; Dewey G. Cornell, University of Virginia

Assessment of school climate has taken on national importance. We investigate agreement among reports obtained by students and other adults in the school through a multilevel CT-C(M-1) analysis of 294 high schools. Focus is on true-score trait variance that was shared across measures of the same trait obtained by different informant types.

Detecting Different Rater Effects Using Trend Scoring Statistics
Widad Abdalla, The University of Iowa; John R. Donoghue, Educational Testing Service; Deborah Harris, University of Iowa

In trend scoring, a set of responses from Time A are rescored by Time B raters. This is used to monitor for rater drift when administering CR items on multiple occasions. The purpose of this study is to detect rater effects in the context of trend scoring using trend-monitoring statistics.

Simulation-Based Investigation of Optimal Modeling Approaches for Structural Equation Models With Ordinal Variables
KwangheeJung, TexasTechUniversity; JaehoonLee, TexasTechUniversity; SeungmanKim, TexasTechUniversity; Heungsun Hwang, McGill University

A Monte Carlo simulation study was designed to compare two alternative approaches to structural equation modeling—generalized structured component analysis with uniqueness terms for accommodating measurement error (the ALS estimator) vs. covariance structure analysis (the ML or the WLSMV estimators)—in terms of parameter recovery with ordinal variables

The Identification of Latent Class Membership in the Mixture Rasch Model
Tongyun Li, Educational Testing Service; Ming Li, Georgetown University; George Macready, University of Maryland

This paper is an investigation of the accuracy of latent class assignment in the MRM. A simulation is conducted in which the latent class assignment is evaluated by the estimated and actual sample proportion of correctly classified respondents, and relative entropy. Additionally, a relative index of correct classification is proposed.

Using the Iterative Latent Class Analysis Approach to Improve Attribute Accuracy

Zhehan Jiang, The University of Alabama - Tuscaloosa

This paper proposes an iterative latent class analysis (ILCA) approach for estimating attributes in cognitive diagnostic modeling. The ILCA is constructed on the expectation maximization algorithm implemented on latent class estimation. Simulation shows the ILCA outperforms its competitors in many conditions.

An Exploration of Considerations for Fitting Dynamic Bayesian Networks With Latent Variables

Ray E. Reichenberg, The George Washington University

The results of a simulation study examining considerations (e.g., sample size, measurement quality) for fitting dynamic Bayesian networks (DBNs) containing latent variables are presented. This evaluation was framed in terms of parameter recovery which was assessed via multiple indices. Recommendations are offered for practitioners wishing to employ DBNs.

Using the Proportion of Flagged Items by S-X^2 for Checking Item Response Theory Assumptions

Jie Xu, Florida State University; Insu Paek, Florida State University; Yan Xia, University of Illinois, Urbana-Champaign; Ki Matlock Cole, Oklahoma State University

No study exists in the literature which assesses the utility of the proportion of flagged items in a test by an item fit index to detect violations of unidimensional IRT assumptions. This study examined the usefulness of this approach using the S-X2 item fit index under uni- and multidimensional data.

Minimizing Classification Errors With Unknown True Cut Scores and Software for Standard-Setting

Jesse Pace, University of Kansas; Irina Grabovsky, National Board of Medical Examiners

We develop a method for finding an optimal cut-score for Pass/Fail examinations which incorporates uncertainty about the'true' point separating proficient examinees from non-proficient ones. We derive false positive and false negative probabilities, introduce several classification metrics, and present software we have developed which performs these calculations for the user.

Effect of Reducing the Number of Distractors in Multiple-Choice Items

Yu Zhang, The Federation of State Boards of Physical Therapy; Aijun Wang; Lorin Mueller, Federation of State Boards of Physical Therapy

Distractors in a multiple-choice (MC) item play a critical role in determining the item's psychometric properties. Researchers have proposed the optimal number of distractors. This study examines the effect of reducing the number of distractors in MC items on the scores of a criterion-referenced test. The preliminary findings indicate that reducing the number of distractors inflates the test score for some candidates.

Prerequisite Structure Finding Using the Conjunctive Root Causes Model

Xinchu Zhao, University of South Carolina; Benjamin Deonovic, ACT, Inc.; Gunter Maris, ACTNext

This study proposes a novel network psychometric model, the root causes model, for finding the prerequisite structure between items and/or skills. The model is shown to be a generalization of classic cognitive diagnostic models (CDMs). The fraction subtraction data set, is scrutinized using this model.

Multigroup Cognitive Diagnostic Joint Modeling of Responses and Response Time

Hong Jiao, University of Maryland-College Park; Manqian Liao, University of Maryland - College Park; Dandan Liao, AIR; Peida Zhan, Zhejiang Normal University

This study proposes multigroup joint modeling of item responses and response time for cognitive diagnosis to account for the conditional between-subject dependence between responses and response time using a multigroup structure and the conditional dependence of response time on item responses. Both real data and simulation data are analyzed.

Bayesian Applications of Estimated True Score Change Models and Comparison to Student Growth Percentiles

John Denbleyker, Houghton Mifflin Harcourt; Ye Ma

This study introduces the Bayesian version of Lord-McNemar's estimated true gain and compares it to the Kelley's regressed score alternative as well as the observed gain and SGPs. Multiple data sets are used to facilitate the analysis and comparison of these models. An inferential framework for measuring growth is elucidated.

Understanding Repeater Subgroup Differences in Retesting Score and Time Improvement

Jiawen Zhou, Educational Testing Service; Yi Cao, Educational Testing Service; Morgan James, ETS

Many testing programs allow examinees to retest. The purpose of this study is to explore retest effects in terms of test score gains and response time improvements among different repeater groups. The characteristics of repeater groups and how repeaters change their item responses over multiple attempts are also studied.

Rating Scale Design and the Effect of Extreme Response Style

Sien Deng, ACT, Inc.; Daniel M. Bolt, University of Wisconsin - Madison

While item parallelism is often desirable for the rating-scale design, it likely augments the biasing effects of extreme response style (ERS) under repeated measures. This study investigates such consequences using a multilevel MIRT model. The findings suggest that greater item heterogeneity appears to better detect ERS and reduce ERS bias.

Multilevel Validation of Mathematics Scores From a Cluster Randomized Controlled Trial

Robert F. Dedrick, University of South Florida; Doug Rohrer, University of South Florida; Marissa Hartwig, University of South Florida

We used a multilevel framework to evaluate the validity of mathematics scores in a cluster randomized trial (54 classrooms, 787 students). Results supported the two-level structure of the test and measurement invariance across intervention and control conditions. The importance of aligning the validation framework with the research design is discussed.

Approaches to Scoring Multiple-Choice Multiple-Select Items: A Comparison Using NAEP Data

William Lorié, Capital Metrics

Multiple-choice multiple-select (MCMS) test items are designed to support more than one keyed option. National Assessment of Educational Progress piloted MCMS items in its 2016 grade 4 and 8 mathematics digital-based assessments. The effects of scoring these MCMS items using dichotomous, rubric-based, and rule-based methods are presented and discussed.

Parametric and Nonparametric Differential Item Functioning Detection in Cognitively Diagnostic Assessments
Sook Hyun Park, The University of Texas - Austin; Hyeon-Ah Kang, The University of Texas at Austin

The study presents nonparametric approaches for identifying DIF in cognitively diagnostic measurement. Substantive simulation studies and real data analyses are implemented to evaluate the performance of the proposed methods in comparison with parametric methods (e.g., Wald tests). The study provides implications for the use of both procedures in practice.

Developing a Framework for Adapting Narrative Feedback Prompts
Hollis Lai, University of Alberta; Vijay Daniels, University of Alberta

Assessment of medical student's clinical performance rely on written narrative feedback from practicing physicians. Current narrative prompts solicit generic feedback and lack task-specificity. To improve the quality of clinical feedback, we introduce a novel method of adapting the narrative prompts and present a pilot application of the framework.

Exploring Item Scoring Methods for Technology-Enhanced Items in Computerized Adaptive Tests
Shu-Chuan Kao, Pearson VUE; Joe Betts, Pearson VUE; William Joseph Muntean, Pearson; Qian Hong, NCSBN

The use of technology-enhanced items (TEIs) brings both great possibilities for item development and new challenges for item scoring. This study explored different methods to identify the appropriate number of scoring categories for TEIs with the hope to enhance polytomous scoring in the partial credit model.

Evaluating the Effects of Extended Time Accommodation on Writing Performance in NAEP 2011
Youmi Suk, The University of Wisconsin - Madison; Young Yee Kim, American Institutes for Research; Xiaying Zheng, American Institutes for Research

Most testing programs provide extended-time accommodation (ETA) to help certain students to demonstrate their abilities. This paper investigates the effects of ETA on writing performance using NAEP Grade-8 Writing 2017 assessment data. We compare the performance of ETA students with that of non-ETA students using a propensity score matching approach.

Assessing the Reliability of Single-Item Measures
Kimberly F. Colvin, University at Albany - SUNY; Guher Gorgun, University at Albany - SUNY

The concept of reliability in the context of single-item measures is addressed. Several methods for estimating the reliability of a single-item measure will be compared, including one novel approach. The results will be compared for two different single-item measures.

The Evidentiary Value of Process Data for Investigating Test-Taking Processes
Tanesia Beverly, University of Connecticut

Test-wiseness has generally been studied using self-report instrument. This study explores an approach for test-score validation that examines test-takers' strategies for taking a reasoning ability test using process data. Latent class analysis (LCA) is applied to the data to classify students as test-wise or test-naive.

National Council on Measurement in Education Invited Electronic Board 1
Andre A. Rupp, Educational Testing Service (ETS)

Saturday, April 6, 2019
10:25 – 11:55am, Manitoba, Coordinated Session

Blending Evidence-Centered Design and Universal Design for Learning in Next-Generation Science Assessment

Chair: Meagan Karvonen, The University of Kansas
Discussant: Brian Gong, Center for Assessment

This coordinated session will present implementation and evaluation of activities from a current USED Enhanced Assessment Grant addressing innovations in several aspects of assessment system design in the context of a next generation science assessment based on alternate achievement standards. The transition to more complex science performance expectations increases challenges in developing assessments that engage students' higher-order thinking skills without introducing accessibility barriers. Building on prior work, strategies for incorporating evidence-centered design and universal design for learning into the extension and evaluation of science learning map models, Essential Element concept maps, and innovative test content with simulated science inquiry and student-engagement features will be described. Additionally, results from cognitive labs that explored interactions of students with various innovative test features will be presented along with results of student and teacher interviews. Examples of how student and teacher experiences further shaped test design plans will be also be discussed.

Overview of Next Generation Science Standards–Aligned Learning Map Models and Assessment Design Considerations
Meagan Karvonen, The University of Kansas

Designing and Evaluating Accessible Science Learning Map Models
Lori Andersen, University of Kansas

Assessment Design: Integrating Evidence-Centered Design and Universal Design for Learning
Russell E. Swinburne Romine, The University of Kansas

Using Cognitive Labs to Evaluate Innovative Features of Next Generation Science Standards–Aligned Assessments
Gail C. Tiemann, The University of Kansas

Saturday, April 6, 2019
10:25 – 11:55am, Quebec, Invited Speaker Session

Communicating Your Research to the Media
Chair: Rebecca Zwick, Educational Testing Service
Panelist: R. Holly Yettick, Editorial Projects in Education, Inc.
Panelist: Emily Richmond, Education Writers Association

This special Presidential session will feature education reporters Emily Richmond, the public editor of the Education Writers Association, and Holly Yettick, the director of the Education Week Research Center. They will offer their advice on communicating research findings to the media from the perspective of those who do the day-to-day work of sifting through research and pitches. This will be an interactive session, in which the reporters will provide hands-on exercises for the attendees to help them understand how a piece of research makes its way from a press release into a published story. There will be 30 minutes reserved for Q&A. The session will be chaired by Rebecca Zwick.

Saturday, April 6, 2019
10:25 – 11:55am, Salon A, Paper Session

Computerized Adaptive Testing: New Directions and Opportunities

Discussant: Walter D. Way, The College Board

New Conditional Measures of the Amount of the Adaptation for Adaptive Tests

Unhee Ju, Michigan State University; Mark D. Reckase, Michigan State University

Existing indices of the amount of adaptation are useful, but they are limited to evaluating adaptivity over groups of examinees, rather than individuals. This study proposes four measures of how much adaptation occurs conditional on profi y level and examines their performance for item pool compositions and test designs through simulations.

On-the-Fly Estimation of Person and Item Parameters in Adaptive Testing

Shengyu Jiang, University of Minnesota - Twin Cities; Chun Wang, University of Washington

We proposed an on-the-fly Bayesian algorithm in computerized adaptive testing (CAT). The algorithm can estimate item and person parameters simultaneously without assuming a pre-calibrated item bank. Through a simulation study, we demonstrated that this algorithm could be successfully adapted to CAT with good parameter recovery.

Comparing Multidimensional Computerized Adaptive Testing to Adaptive Test Batteries for Within-Subject Multidimensional Assessments

Tyler Matta, Pearson; Kirk A. Becker, Pearson; David Shin, Pearson; Xinrui Wang, Pearson VUE

Navigation through a simple structure multidimensional adaptive test can be done in two ways: multidimensional CAT or as an adaptive test battery. This paper focuses on the psychometric considerations of the two approaches. Specifically, bias, mean-square error, and number of items administered for each domain are compared.

Computerized Adaptive Testing Algorithms (Marginal, Joint, Iterative) for Student Evaluation of Teaching

Chia-Wen Chen, The Education University of Hong Kong; Chen-Wei Liu, Chinese University of Hong Kong; Ming Ming Chiu, The Education University of Hong Kong; Wen-Chung Wang, University of Hong Kong, Hong Kong

No one has applied Computerized Adaptive Testing (CAT) to Student Evaluation of Teaching (SET). We introduce three algorithms for CAT of SET (Marginal, Joint, Iterative) and evaluate them via simulations. The Iterative method showed the highest reliability, lowest RMSE, and little bias. Joint or Marginal were worse. Random was the worst.

Using Off-Grade Items in Adaptive Testing: When Is It Appropriate?

Shuqin Tao, Curriculum Associates; Daniel F. Mix, Curriculum Associates; James Cunningham, University of North Carolina - Chapel Hill

This study is intended to investigate the underlying causes for differential item functioning exhibited by items when administered off-grade in adaptive testing. Data came from an adaptive assessment administered to school districts nationwide. Insights gained will help inform item selection strategies in adaptive algorithms to select appropriate off-grade items.

Saturday, April 6, 2019
10:25 – 11:55am, Salon B, Paper Session

Emerging Research in Linking and Equating

Discussant: Jinghua Liu, The Enrollment Management Association

Quantifying True Score Bias in Nonlinear Equating and Score Transformation

Matthias Von Davier, National Board of Medical Examiners

This paper explores effects of using non-linear transformations on not perfectly reliable test scores. A requirement of any valid scale transformation can be said to be that any test takers'true score on the new test should be mapped onto the true score of the reference form. For linear transformations, the true score on the new scale, the target of the transformed observed scores, equals the transformed true score of the untransformed score. This does not hold for non-linear transformations if tests are not perfectly reliable. An approximation of this bias in conditional expectations of transformed observed scores is provided and illustrations will be presented.

Evaluating Weighting Methods for Linking Assessment With No Anchor Items

Shuhong Li, Educational Testing Service; Jiahe Qian, ETS

A recently proposed weighted linking method using minimum discriminant information is compared with the poststratification equipercentile equating and the unweighted nonparametric equipercentile equating in linking assessments with no anchor items. RMSE is used as the evaluation criterion. Real data from a large-scale assessment with multiple forms are used.

The Impact of Estimation Procedures for Latent-Variable Distributions on Item Response Theory Linking

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

The purpose of this study is to compare the relative performance of separate, concurrent, and fixed parameter calibration using three different estimation procedures for the latent-variable distributions. Comparison will be made through simulation for the unidimensional three-parameter logistic model and the multidimensional bifactor model under various population distributions.

Asymptotic Standard Errors of Polytomous GPCM True Score Equating by Response Function Equating Methods

Zhonghua Zhang, University of Melbourne

In this study, building on the works by Ogasawara (2000, 2001a, 2001b) and Wong (2015), formulas are derived for applying delta method to estimate the standard errors of item response theory (IRT) true score equating for polytomous generalized partial credit model (GPCM, Muraki, 1992) using the item and test response function equating methods, in the context of the common-item nonequivalent groups equating design. Simulation study is further conducted to compare the results derived from the delta method with those produced by the bootstrap method.

Saturday, April 6, 2019
10:25 – 11:55am, Territories, Coordinated Session

Practical Measurement for Improvement Science: Principles and Applications

Chair: Andrew Ho, Harvard University
Discussant: Anthony S. Bryk, The Carnegie Foundation for the Advancement of Teaching

What is Improvement Science, and how are practitioners of this growing discipline designing and evaluating their measurement tools? Improvement Science is a method that users employ to identify, understand, and solve specific problems of practice (Bryk et al., 2015; Langley et al., 2009). The method requires measurements that support rapid cycles of disciplined inquiry, to detect the effects of prototyped changes. Whereas large-scale tests emphasize standardization and precision, measures for Improvement Science must typically be more convenient for local practitioners to develop, administer, score, interpret, and use. Paper 1 introduces principles for practical measurement for improvement. Papers 2, 3, and 4 present examples of tools, for improving writing, mathematics, and on-time graduation, respectively. The chair, presenters, and the discussant will also identify productive tensions between modern testing standards and practical measurement for improvement. Are traditional metrics for reliability and precision relevant for measures whose purpose is to detect the effects of a prototyped change? And how should these practical measurement tools quantify growth? This symposium integrates the ideas and memberships of two organizations deeply committed to improving the use of measurement tools in education: The National Council on Measurement in Education and the Carnegie Foundation for the Advancement of Teaching.

Practical Measurement for Improvement: An Introduction
Jon Norman, Carnegie Foundation for the Advancement of Teaching; Sola Takahashi, WestEd

Developing Practical Measures to Inform Instructional Improvement Initiatives in Mathematics
Paul A. Cobb, Vanderbilt University; Kara J. Jackson, University of Washington - Seattle; Marsha M. Ing, University of California - Riverside

The Using Sources Tool: A Practical Measure for Identifying Next Instructional Steps
Linda D. Friedrich, National Writing Project

Using the Freshman On-Track Indicator to Improve High Schools in Chicago
Elaine M. Allensworth, University of Chicago

Saturday, April 6, 2019
12:20 – 1:50pm, Alberta, Invited Speaker Session

Classroom Assessment and Educational Measurement

Chair: Susan M. Brookhart, Duquesne University
Discussant: James H. McMillan, Virginia Commonwealth University

This session highlights five chapters from an upcoming book in the NCME Book Series, Classroom Assessment and Educational Measurement. The purpose of the book is to investigate how classroom assessment perspectives can inform educational measurement and how educational measurement perspectives can inform classroom assessment. The issues raised in the five chapters in this session can be summarized as a series of questions: What is classroom assessment? What does it mean for classroom assessment information to be valid, trustworthy, and fair? Can advances in digital technology enhance classroom assessment? The first paper shows how a functional perspective on validity is more salient in classroom assessment than a measurement perspective. The second paper discusses the trustworthiness of classroom assessment information, framed from a more practitioner-oriented perspective. The third paper describes how the fairness standards from the Standards for Educational and Psychological Testing can help address issues of diversity, equity, and inclusion in classroom assessment. The fourth paper describes an approach to creating a system of embedded assessment based on learning progressions that can be used to monitor and track growth in student understanding over time. The fifth paper describes how digital technologies can support and advance classroom assessment practices.

Fairness in Classroom Assessment
Joan L. Herman, University of California - Los Angeles; Linda L. Cook, Educational Testing Service

Digital Technologies Supporting and Advancing Assessment Practices in the Classroom
Mike K. Russell, Boston College

Defining Trustworthiness for Teachers' Multiple Uses of Classroom Assessment Results
Alicia C. Alonzo, Michigan State University

Perspectives on the Validity of Classroom Assessments
Michael T. Kane, ETS; Saskia Wools, Cito

Learning Progressions and Embedded Assessment
Derek C Briggs, University of Colorado - Boulder; Erin Marie Furtak, University of Colorado - Boulder

Saturday, April 6, 2019
12:20–1:50pm, Algonquin, Coordinated Session

Scales and Norms for Achievement and Growth: Approaches and Applications

Chair: Yeow Thum, Northwest Evaluation Association
Discussant: Mark Reckase

Under the Every Student Succeeds Act (ESSA) of 2015, school are increasingly being held accountable for student growth, rather than just static achievement. To support inferences about growth, we need to develop growth scales and, with that, build norms for growth. While most assessments provide norm-referenced scores (percentiles) relative to a target examinee reference population, norms for growth patterns has not been widely implemented in large-scale assessments. This symposium presents new efforts to build growth scales from fitting growth curves to longitudinal achievement data. After a suitable growth curve is developed, the predictive distribution of scores for a given point in time defines the corresponding conditional norms for achievement. Such relationships between achievement and time provide the basis for developing normative growth scales. We describe the development of growth norms for a widely-used interim assessment and use these norms to benchmark typical patterns of growth across grades. Additionally, we discuss approaches for norming educational data when time is measured on a continuous scale, whether in years or instructional weeks

Developing Performance and Growth Norms for Students and Schools
Yeow Thum, Northwest Evaluation Association; Megan Kuhfeld, NWEA

Visualizing Student and School Achievement and Growth Using Shiny
Megan Kuhfeld, NWEA; Yeow Thum, Northwest Evaluation Association

Empirical Benchmarks From Growth Norms: School ICCs for Gender Gaps and Summer Loss
James Soland, NWEA; Yeow Thum, Northwest Evaluation Association

The Application of Continuous Test Norming With Generalized Additive Models for Location, Scale, and Shape
Lieke Voncken, University of Groningen; Casper Albers, University of Groningen; Theo van Batenburg, University of Groningen; Marieke Timmerman, University of Groningen

Saturday, April 6, 2019
12:20 – 1:50pm, Ballroom, Coordinated Session

Automated Assessment of Scientific Reasoning: Developments in the Field

Chair: Jonathan F. Osborne, Stanford University
Discussant: James W. Pellegrino, University of Illinois at Chicago

This symposium will be an opportunity to hear about 4 research-based approaches to the computer-based assessment of scientific thinking and the nature and quality of the feedback these approaches provide. Machine learning and AI are set to transform complex tasks in the coming decades and this symposium will present developments in the field of student assessment in science. Three of the four approaches draw on machine learning and natural language processing to automate the process of scoring. The fourth examines whether it is possible to assess higher order reasoning using forced choice and selected responses. Such approaches have the potential to provide a) immediate student feedback, and b) immediate data for the teacher to inform the pedagogic choices that they then make. Thus, they have the potential to enhance the quality of formative assessment and improve personalized feedback. The four papers in this coordinated session use data drawn from Grade 6 to undergraduate students and a diversity of scientific contexts. The symposium provides an opportunity to learn about both the success and challenges of these different approaches and to contrast the strengths and weaknesses of each.

Formative Assessment of Scientific Argumentation Practice Enabled by Automated Text Scoring
Margarita Olivera Aguilar, Educational Testing Service; Hee-Sun Lee, Concord Consortium; Ou Lydia Liu, ETS; Amy R. Pallant, Concord Consortium

Assessing Higher Order Reasoning Using Technology-Enhanced Forced-Choice Item Types in the Context of Science
Mark R. Wilson, University of California - Berkeley; Linda Morell, University of California - Berkeley; Jonathan F. Osborne, Stanford University; Sara Dozier, Stanford University; Weeraphat Suksiri, University of California - Berkeley

Characterizing the Composition, Structure, and Coherence of Students' Evolutionary Explanations Using Automated Text Scoring
Ross H. Nehm, Stony Brook University of New York At Stony Brook; Minsu Ha Kangwon, Stony Brook University - SUNY

Using Automated Analysis to Assess Middle School Students Competence With Scientific Argumentation.
Kevin Haudek, Michigan State University; Jonathan F. Osborne, Stanford University; Christopher D. Wilson, Biological Sciences Curriculum Study

Saturday, April 6, 2019
12:20 – 1:50pm, British Columbia, Paper Session

Research Advancing Item Calibration Methods

Discussant: Tracy Gardner, New Meridian Corporation

Improving the Fixed Item Parameter Calibration With the Fixed Prior

Sung-Hyuck Lee; JP Kim, ETS

The prior is important with the fixed item parameter calibration (FIPC) where the prior is updated during the EM cycles. When the prior is inaccurately updated, it adversely affects the item parameter estimation with the FIPC. In this study, the fixed prior calibration method is proposed to improve the FIPC.

Using Item Content and Category Information to Predict Item Parameters

Kristin M. Morrison, ACT, Inc.; William Skorupski, Amira Learning; David Carmody, ACT, Inc.; Richard Meisner, ACT, Inc.; Justin Paulsen, Indiana University

This study used a variety of modern techniques to create items from task models, after which items were field tested. Coded features of items and item categories were then used as independent variables in hierarchical item response models to enhance calibration precision and explain the variance in item parameter estimates.

Contextual Effects of Response Omission Corrections

Michael Chajewski, Kaplan Test Prep

The herein delineated work demonstrates that contextual differences likely require differential treatment of omitted responses. Examinees who desire to receive better formative insight should be instructed on the effects of omitting responses as part of communicating educational measurement attributes and their immediate relevance to constituents.

Scoring and Calibration for EBSR Items

Dong-In Kim, DRC; Christie L. Plackner, DRC; Vince Struthers, DRC; Mayuko Simon, DRC

In large scale assessments, a polytomous score based on conditional scoring for two EBSR MC parts is often used to IRT calibration. However, the calibration with conditional scoring approach sometimes causes item fit issues due to guessing for one score point. This study examines six methods to resolve this issue.

Saturday, April 6, 2019
12:20 – 1:50pm, Imperial Room, Graduate Student Poster Session

GSIC Graduate Student Poster Session 1

Bayesian Model Selection in Mixture Item Response Theory and IRTree Model for Response Style

Juyeon Lee, University of Georgia - Athens; Allan S. Cohen, University of Georgia

In this study, we examine the usefulness of PPMC as a Bayesian model selection index for aberrant response styles. Mixture IRT and IRTree models will be compared using the science attitude scale of PISA 2015. The results will be amplified with graphical presentations of posterior predictive values.

Differential Item Functioning Detection Procedure for Multistage Computer Adaptive Tests

Christiana Aikenosi Akande, University of Florida; Meng Wu, ETS; Xueli Xu, Educational Testing Service

This study examines the validity of utilizing the current NAEP DIF program for a multistage test design. Data was simulated using operational NAEP parameters for Linear and MST administrations. The current NAEP DIF program utilizes the Mantel-Heanszel procedure. Results indicate that the program produces biased estimates for the MST design.

Investigating Rater Behavior on an Objective Standard-Setting Exercise

Karen Fong, American Society for Clinical Pathology - Board of Certification

This study investigated ratings from 28 raters on an Objective Standard Setting exercise for a medical certification. Instead of having all 28 raters provide ratings for all items, the items were separated by content area and raters rated items according to their expertise.

Using Person-Fit Statistics to Detect Response Styles

Yingbin Zhang, University of Illinois, Urbana-Champaign; Tuo Liu, Tianjin Normal University

This study attempted to use four person-fit statistics, including Gp, Gp-n, U3p and lz-p , to detect response styles. Results showed that they could detect extreme and disacquiescence response style, but the detection rates for disacquiescence response style were relative low. They were unable to detect acquiescence and midpoint response style.

Measurement Invariance of the Kutcher Adolescent Depression Scale Across Gender and Marital Status

Mahnaz Shojaee, University of Alberta; Amin Mousavi, University of Saskatchewan; Mehrdad Shahidi, Mount Saint Vincent University; Ying Cui, University of Alberta

Measurement equivalence is one of the most important aspects of fairness in testing, and differential item functioning (DIF) is one of the techniques to assess it. Evaluating DIF in KADS-11 self-reported diagnostic depression instrument by means of Ordinal Logistic Regression, we found some item exhibiting DIF but with negligible magnitude.

Comparison of the Methods to Detect the Noninvariant Items in Multigroup Confirmatory Factor Analysis

Minju Hong, University of Georgia - Athens

This study examined the performance of the methods to detect non-invariant items. Factor ratios test, backward MI method, and forward CI methods were tested by manipulating the conditions of sample size. In unbalanced and small sample conditions, except for factor ratio test method, other two methods outperformed in the detection.

A Comparative Study of Assessing Dimensionality of Item Response Theory Models

Wenjing Guo, The University of Alabama; Youn-Jeng Choi, The University of Alabama

This study compared the accuracy of traditional and revised parallel analyses regarding to determine the number of underlying factors in IRT. Preliminary results showed (1) when unidimensional IRT models were generated, traditional parallel analyses performed better; (2) when generated model was multidimensional IRT, revised parallel analyses were generally more accurate.

Improving Item Pool Utilization for Shadow Computerized Adaptive Testing in Variable Length

Hwanggyu Lim, University of Massachusetts; Qi Diao, ETS

Diao and Ren (2018) suggested a way of constructing a Shadow CAT in variable-length. However, the Shadow CAT showed poor utilization of item pool although it performed well in terms of measurement precision. Thus, this study suggests new approaches to improve the pool utilization of the Shadow CAT in variable-length.

Differential Item Functioning in Math Items Generated by an Automatic Item Generator

Eunbee Kim, Georgia institute of technology; Susan Embretson, Georgia Institute of Technology

The purpose of this paper is to test the equivalency of item parameters (DIF) across ethnicity and gender for the items from an automatic item generator. The proportion of items with significant DIF is less than 10%, which further supports generated items as not requiring the traditional extensive tryout procedures.

Examining for Differential Item Functioning in Test of Mathematical Abilities (TOMA-3)

Soyoung Park, The University of Texas - Austin

This study aimed to investigate potential item bias related to gender, race and ethnicity on the Test of Mathematical Abilities--Third Edition (TOMA-3). Differential Item Functioning (DIF) analysis across age, gender, race and ethnicity on an assessment can be considered as a core element of the item bias.

The Performance of Recursive Partitioning Methods for Dealing With Missing Responses

Jiaying Xiao, University of Alberta; Okan Bulut, University of Alberta

This simulation study compared the performances of recursive partitioning methods (CART and Random Forest) with traditional imputation methods (FIML and zero-replacement) for replacing missing responses. The methods were compared based on the accuracy of ability estimates. Results indicated that CART, Random Forest, and FIML performed similarly; zero-replacement was the worst.

Investigating Item Position and Psychological Factor Effects With Missing Data

Nayeon Yoo, Teachers College, Columbia University; Ummugul Bezirhan, Teachers College, Columbia University; Young-Sun Lee, Columbia University

The purpose of this study is to investigate the effects of item position and psychological factors with missing data. Real-world data analyses were conducted using TIMSS 2015 Grade 8 Mathematics data, and simulation studies were conducted to investigate the item position effect considering psychological factors when missing data is present.

Module Assembly and Routing of Cognitive Diagnostic Multistage Adaptive Test

Manqian Liao, University of Maryland - College Park; Hong Jiao, University of Maryland-College Park

This study proposes a procedure for assembling and administering Cognitive diagnostic multistage testing (CD-MST), which strikes a balance between linear diagnostic test and CD-CAT. It reduces the number of required modules, making the development of CD-MST more cost-efficient and feasible. Several module assembly methods and routing strategies were compared.

Evaluation of the Utility of Informative Priors in Structural Equation Modeling With Small Samples

Hao Ma, Southern Methodist University; Akihito Kamata, Southern Methodist University; Yusuf Kara, Southern Methodist University

This study evaluates the performance of different estimators on factor loadings and structural coefficients in terms of bias, RMSE, and SE for factor analysis and SEM models under the ML and Bayesian framework with small sample settings. Simulation conditions varied in sample sizes, mean factor loading, priors, and estimators.

Is It Necessary to Incorporate Response Time to Reduce Speededness Effects?

Lu Wang, The University of Iowa; Robert D. Ankenmann, University of Iowa

This study is intended to explore whether it is necessary to incorporate item response time (RT) information in calibration to deal with speededness effects. When the RT information is not available, can methods that only utilize response patterns to deal with speededness effects provide comparable parameter estimates.

Fast-and-Accurate Effect as a Function of Item Difficulty, Ability, and Remaining Time

Yang Shi, University of California - Berkeley; Paul De Boeck, The Ohio State University; Kyung (Chris) T. Han, The Graduate Management Admission Council

This study aims to explore the fact-and-accurate effect in computerized adaptive testing. The results show the existence of local dependencies, which indicates the hierarchical model and measurement invariance as a function of speed are violated. This violation implies that fast responses do not have the same meaning as slow responses.

A Study of a Fit Index for Explanatory Item Response Theory Models

Heather Anne Handy, Georgia Institute of Technology; Susan Embretson, Georgia Institute of Technology

Applying explanatory item response theory (IRT) models is advantageous when designing and selecting items. A simulation study was conducted to compare an explanatory IRT fit statistic to traditionally used fit indices for assessing model quality. Simulation conditions include varying test length, item diffi y and the number of predictors.

Adaptive Angoff Standard-Setting Method: An Exploratory Study

Ella Gift Banda, University of Massachusetts - Amherst; Joshua T. Goodman, National Commission on Certifi ation of Physician Assistants; John Weir, NCCPA

Standard setting is a critical but time consuming and expensive process. The decisions emanating from the process impacts the validity inferences made from the scores. This study investigates an adaptive variation on the Angoff method that could save time and money, and perhaps even gain efficiency

Class-Specific Responses Processes in Concept Learning and Intelligence Tests

Clifford E Hauenstein, Georgia Tech; Susan Embretson, Georgia Institute of Technology

Concept learning is an important aspect of intelligence tests used to assess educational potential. Item difficulty modeling of two latent classes of examinees on the Woodcock-Johnson III test found class differences in approaches to item solving. Class performances were either limited by rule type or Boolean complexity.

Exploratory Analysis of Process Data to Investigate Students' Learning Outcomes

Yawei Shen, The University of Georgia; Shiyu Wang, University of Georgia

A series of exploratory approaches were used to investigate students' learning behaviors in a learning program using multivariate response data and two types of timing data. Several clusters characterized by students'response accuracy, speediness and engagement were discovered. The findings provide feedbacks about students' learning outcomes from multiple perspectives.

A Critique of Predictive Methods in Evidence-Based Standard-Setting

Yi Xe Thng, Harvard University; Andrew Ho, Harvard University

We critique predictive methods in evidence-based standard setting (McClarty, Way, Porter, Beimers, & Miles, 2013). We demonstrate how cut scores set using predictive methods are just as subject to judgment as traditional standard setting. Additionally, they confound the stringency of the performance standard with the correlation between predictor and outcome.

Comparing Multiple Disability Groups in Alternate Assessments Using Differential Item Functioning

Mahmut Gundogdu, University of California Riverside; Fusun Sahin, American Institues for Research

A large-scale alternate assessment test administered to three different disability groups: Autistics(AUT), Intellectual disabilities(ID), and Multiple Disabilities(MD) were analyzed for DIF using generalized logistic regression. DIF was suitable for comparing ID and MD groups simultaneously relative to the AUT group.

Effects of Probability Threshold Choice on an Adjustment for Guessing With Rasch Modeling

Tom Waterbury, James Madison University; Christine Demars, James Madison University

We investigated a strategy for accounting for guessing with the Rasch model. It involves converting to missing data any item/person encounters below a probability threshold and running the Rasch analysis without that data. We showed the value of the probability threshold has implications for the accuracy of item difficulty estimates.

Using HGLM to Estimate Variance Components Under the Framework of Generalizability Theory

Mingfeng Xue, Beijing Normal University; Ping Chen, Beijing Normal University

This study proposes using hierarchical generalized linear models (HGLM) to estimate variance components under the framework of generalizability theory, and then fully compares it with traditional linear mixed model. The results indicate that HGLM well reflects the variance of the data and provides better estimates of G and D coefficients.

Saturday, April 6, 2019
12:20 – 1:50pm, Manitoba, Paper Session

Fairness Issues in Test Construction

Discussant: Melissa L. Gholson, Educational Testing Service

Assessing Differential Item Response Time: Rationale, Methodology, and Implications

Min Wang, ACT; Mengyao Zhang, National Conference of Bar Examiners; Xiaohong Gao, ACT, Inc.

Assessing differential item response time (DIRT) could inform test development and support test validity and fairness. Conceptualizing DIRT in a spirit similar to differential item functioning (DIF), this empirical study describes varying DIRT patterns, investigates methods for detecting DIRT, and explores the utilization of DIRT in the evaluation of DIF.

Test Construction and Selection Bias: An Investigation Using the Rasch Model

Andrew Jones, American Board of Surgery; Jason P. Kopp, American Board of Surgery; Thai Ong, James Madison University

Selection bias, wherein the use of test scores for classifi tion results in non-equivalent classifi tion errors between subgroups, has received little attention in the psychometric literature. This study uses the Rasch model to demonstrate that the location of maximum test information affects classification errors differentially between subgroups.

Length of Practice and Performance on a Computer-Based English Language Proficiency Screener

Nami Shin, University of California - Los Angeles; Mark Hansen, University of California - Los Angeles; Eunhee Keum, University of California - Los Angeles

This study examines to examine the relationship between length of practice session and students' performance on a computer-based English Language Proficiency (ELP) screening test. Determining an optimal length for the practice session will ensure efficient delivery of the test and support valid inferences about students' language skills.

Selecting Optimal Stopping Rules for a K–12 English Language Proficiency Screening Test

Eunhee Keum, University of California - Los Angeles; Mark Hansen, University of California - Los Angeles; Michelle McCoy, University of California - Los Angeles; Eric Setoguchi, University of California - Los Angeles

English language proficiency (ELP) screening tests must support high stakes placement decisions despite being administered under severe time constraints. This study proposes a simulation-based approach for exploring tradeoffs between classification accuracy and testing time in order to identify optimal stopping rules.

Technology-Enhanced Items: Gender and Ethnicity Differences on Response Time, Test Engagement, and Achievement

Wei He, NWEA

The proposed study will use test events from 4th and 8th graders to examine the impacts of technology-enhanced items on test-taking subgroups in terms of response time, test engagement, and achievement gap. A series of multivariate multilevel models will be constructed to address the research questions.

Saturday, April 6, 2019
12:20 – 1:50pm, Quebec, Invited Speaker Session

Communicating/Depicting Results in Easily Accessible Ways, Across Broad Audiences

Chair: Brent Bridgeman, ETS

Sophisticated psychometrics are of no value if study results cannot be meaningfully described to the relevant audiences. This symposium addresses problems and promising practices in effectively communicating educational test results. One problem addressed is that test users often want information that goes beyond total test scores; specifically, they want subscores from tests that were never designed to yield subscores. It is important to clearly explain to these users what valid inferences can be made from subscores. Second, communicating meaningful indicators of learning progress to teachers and parents who have limited or no understanding of what educational assessments can and cannot do is especially challenging. Third, test critics who look at the ability of admissions tests to predict meaningful outcomes often conclude that the tests must of very limited value based on low correlations that explain very little of the variance in the outcome measures. But looking at the same results in a different way can demonstrate considerable value in the assessments. The symposium will conclude with the perspective of a writer from outside of the psychometric community who effectively communicates the results from educational tests to lay audiences.

Using Subscores to Support Valid Inferences
Richard A. Feinberg, National Board of Medical Examiners

Beyond Correlations: Making Predictive Validity Results Understandable
Brent Bridgeman, ETS

Helping Students and Their Parents Understand the Substance of Learning Progress
Lorrie A. Shepard, University of Colorado - Boulder

Saturday, April 6, 2019
12:20 – 1:50pm, Salon A, Paper Session

Collecting and Communicating Validity Evidence

Discussant: Brian Gong, Center for Assessment

Validity and Reliability of the DAACS Writing Assessment

Diana Akhmedjanova, University at Albany - SUNY; Angela M. Lui, University at Albany – SUNY; Heidi L. Andrade, University at Albany - SUNY; Jason Bryer, Excelsior College

The DAACS system assesses incoming college students' readiness in reading, writing, mathematics, and self-regulated learning. The purpose of this study is to examine validity and reliability evidence for the internal structure of the writing assessment. The evidence suggests that our conceptual framework holds for both human and machine scored essays.

Evidence of validity for a middle grades mathematics universal screener based on relationships to state test scores

Elizabeth Lynn Adams, Southern Methodist University; Leanne R. Ketterlin-Geller, Southern Methodist University

This study examines the validity of a universal screener for predicting performance on a state test. Utilizing hierarchical linear modeling, results show the scores on the universal screener across three timepoints are positively related to test scores ($p < .001$). The results persist after controlling for student and school differences.

Highlighting Actual Interpretations and Uses in Validity Evidence

Marsha M. Ing, University of California - Riverside; Starlie Chinen, University of Washington - Seattle; Kara J. Jackson, University of Washington - Seattle; Thomas M. Smith, University of California - Riverside

This presentation describes evidence for the actual interpretation and use arguments of indicators designed to support instructional improvement efforts around the quality of middle grades mathematics discourse. Findings highlight the need for describing the characteristics of the users and the context in which the practical indicator was used.

Simplifying and Communicating Validity and Reliability Studies for Small "N" Professional Programs

Colleen Thornton MacKinnon, Independent Consultant; Mary Yakimowski, Samford University, Alabama

Educator preparation programs using locally created assessments with small sample sizes encounter challenges when investigating technical properties as required by accreditation standards. In this case study, we examine reliability and validity with same size, including pathways for continuous improvements while furthering the abilities of stakeholders to investigate technical properties.

Examining the Alignment of Assessment Literacy Curriculum With the Classroom Assessment Standards.

Aarti P. Bellara, University of Connecticut - Storrs

The purpose of this study was to examine the alignment of the curriculum for a preservice teacher education assessment course with the Classroom Assessment Standards to identify which standards were explicitly addressed. Findings support the need for teacher educators to continue to examine ways to prepare assessment literate educators.

Saturday, April 6, 2019
12:20 – 1:50pm, Salon B, Paper Session

Cultural Considerations in Test Development and Validity

Discussant: Bruno D. Zumbo, The University of British Columbia

Development and Validation of Thai Versions of the International Personality Item Pool

Chakadee Waiyavutti, University of Iowa

In this study, we describe translation and validation of the IPIP, Big-Five inventory into the Thai language. Results showed that the customized Thai IPIP yielded psychometric properties as good or better than the original versions in both Thai and American samples.

Extending the Cultural Fairness Review Process in the Ability Testing Context

Joni M. Lakin, Auburn University

This paper describes the development of a picture-based K-12 cognitive abilities test designed to minimize the infl e of culture-specifi knowledge. We will discuss cross-cultural test adaptation practices and our development of resources for cultural fairness reviews that explicitly address cultural loading of pictures and concepts beyond the usual achievement domains.

Modeling Choices of PISA Scaled Scores: An Investigation Using Multilevel Item Response Theory Framework

Jayashri Srinivasan, University of California - Los Angeles

Increasingly, PISA tests are influencing education policies across multiple countries. This study makes use of the PISA 2009 data and a multilevel IRT framework to examine the issue of measurement invariance and modeling choices to investigate student's access to various instructional practices across rural and urban regions.

Issues for Cultural Validity in Science Contextualized Items

Nixi Wang, University of Washington - Seattle; Min Li, University of Washington; Dongsheng Dong, University of Washington - Seattle

Unintended testing consequences including its social and cultural implications have drawn increasing attention in recent years. In this paper, we propose a conceptual framework for evaluating the cultural validity of contextualized items and provide empirical evidence by analyzing a pool of released PISA items based on expert reviews.

Saturday, April 6, 2019
12:20 – 1:50pm, Territories, Coordinated Session

The Assessment of English Writing Skills in Secondary Education in Europe

Chair: Stefan Daniel Keller, University of Applied Sciences of Northwestern Switzerland
Discussant: Andre A. Rupp, Educational Testing Service (ETS)

The symposium addresses the assessment of English as a foreign language (EFL) writing proficiency in Europe. So far, a comprehensive investigation of this important area of foreign language proficiency has been neglected in this research area. This is surprising considering the practical importance of written text, especially in view of the large amount of high-stakes written exams in EFL education in secondary school and standardized second language tests. This symposium aims to address this research gap by presenting fresh and ground-breaking empirical research on the assessment of complex productive EFL writing skills at upper-secondary level. We consider issues related to text assessment from different perspectives, presenting research on both automated and human scoring, including samples of professionally trained raters, pre- and in-service teachers, as well as writing experts. Further, the impact of different text features on judgment accuracy is analyzed with a focus on the role of text length in association with text quality. Overall, the session aims at giving its audience a comprehensive overview of cutting-edge research perspectives on EFL writing assessment, both in a European and in an international context. Methodological issues associated with both human and automated scoring regarding the assessment of text quality are discussed critically.

Study Context and Overview
Stefan Daniel Keller, University of Applied Sciences of Northwestern Switzerland; Olaf Koeller, Leibniz Institute for Science and Math Education

Score Reporting: Writing Experts' Text Assessment Using the CEFR
Jennifer Meyer, Leibniz Institute for Science and Mathematics Education; Johanna Fleckenstein, Pädagogische Hochschule FHNW; Stefan Daniel Keller, University of Applied Sciences of Northwestern Switzerland; Olaf Koeller, Leibniz Institute for Science and Math Education

Considering Text Characteristics: The Role of Essay Length in Writing Assessment
Anna Lara Paeske, Leibniz Institute for Science and Mathematics Education; Jennifer Meyer, Leibniz Institute for Science and Mathematics Education; Thorben Jansen, University of Kiel; Johanna Fleckenstein, Pädagogische Hochschule FHNW; Stefan Daniel Keller, University of Applied Sciences of Northwestern Switzerland; Olaf Koeller, Leibniz Institute for Science and Math Education

Considering Rater Characteristics: Differences in Judgement Accuracy of Pre- Versus In-Service Teachers
Thorben Jansen, University of Kiel; Cristina Voegelin, University of Applied Sciences of Northwestern Switzerland; Nils Machts, University of Kiel; Jennifer Meyer, Leibniz Institute for Science and Mathematics Education; Stefan Daniel Keller, University of Applied Sciences of Northwestern Switzerland; Olaf Koeller, Leibniz Institute for Science and Math Education; Jens Moeller, University of Kiel

Saturday, April 6, 2019
2:15 – 3:45pm, Alberta, Coordinated Session

Measurement in Adaptive Learning Systems: Challenges and Solutions

Chair: Maria Bolsinova, ACT, Inc.
Discussant: Maria Bolsinova, ACT, Inc.

Adaptive learning systems are designed to dynamically adjust the level or type of practice and instruction material based on an individual learner's abilities. Measurement plays an important role in adaptive learning, since monitoring of the development of learners'skills is crucial to adapt the learning material to their level. However, the adaptive, dynamic and large-scale nature of these learning systems poses challenges for traditional measurement models. To address these challenges, innovative methods and algorithms have been developed within different adaptive learning systems. These solutions are often not limited in application to the particular system, but can be of interest for the general educational measurement community. In this symposium three adaptive learning systems are presented which cover different educational domains (primary school mathematics, preparation for college admission tests, language learning) and tailor to different populations of learners (young children, high-school students, adults). The three first presentations describe these leaning systems, and specifically address the measurement issues they encountered and the solutions that they have developed. The last presentation deals with a particular feature of adaptive learning, which is available in all these three systems, namely the opportunity of learners to ask for hints during their practice, and its consequences for measurement.

An Urn-Scheme to Track Accuracy and Response Time Measures in Adaptive Learning Systems
Abe Hofman, University of Amsterdam; Maria Bolsinova, ACT, Inc.; Han van der Maas, University of Amsterdam; Gunter Maris, ACTNext

Toward Dynamic Adaptation and Personalization in ACT Academy: A Free Online Learning Platform
Benjamin Deonovic, ACT, Inc.; Michael Yudelson, ACTNext; Pravin V Chopade, ACT, Inc.; Steve Polyak, ACT, Inc.

Improving Language Learning With Data: Measuring Learning (and Forgetting) at Scale
Burr Settles, Duolingo; Masato Hagiwara, Duolingo; Bozena Pajak, Duolingo; Joseph Rollinson, Duolingo

Hints in Adaptive Learning Systems: Consequences for Measurement
Maria Bolsinova, ACT, Inc.; Benjamin Deonovic, ACT, Inc.; Meirav Arieli-Attali, ACTNext by ACT, Inc.; Burr Settles, Duolingo; Masato Hagiwara, Duolingo; Alina A. Von Davier, ACT, Inc.; Gunter Maris, ACTNext

Saturday, April 6, 2019
2:15 – 3:45pm, Algonquin, Coordinated Session

Score Reporting in Ongoing Testing Environments: Reporting Challenges and Innovative Solutions

Chair: Suleyman Olgar, Florida Department of Education
Discussant: April L. Zenisky, University of Massachusetts - Amherst

Modern enhancements in technology have allowed for onsite, instant reporting of test results after concluding a test administration session. This coordinated session brings four papers that address critical test score reporting challenges and potential innovative solutions in ongoing educator certifi tion testing programs. Each paper summarizes research findings and possible solutions for reporting measurement precision in individual reports, improving the quality of the reported scores to individuals and institutions, and the quality of the scoring process and examinee feedback for performance assessments. The fi paper researches measurement precision for individual strand/objective reports. The second paper researches score reporting practices for institutions of higher education and training. The third paper discusses bias concerns arising from score verification. The fourth paper presents potential and innovative solutions to score reporting challenges while ensuring test security and integrity on performance assessments. This session presentations discuss innovative solutions to reporting challenges in ongoing educator testing programs and provide recommendations to practitioners to overcome these and similar issues in operational and research settings.

How Consistent Are the Ups and Downs of Individual Strand/Objective Score Profiles?
Alvaro J. Arce, Pearson

Strategies for Reporting Tailored Test Performance Information to Varying Classes of Test Users
Lauren White, Florida Department of Education

Score Verification and Potential Challenges in Certification Testing
Leah Kaira, Pearson

Holistic Scoring and Reporting: Enhancing Reporting While Ensuring Test Security and Integrity
Suleyman Olgar, Florida Department of Education

Saturday, April 6, 2019
2:15 – 3:45pm, Ballroom, Coordinated Session

Advanced Psychometrics for Process Data Analysis in Large-Scale Assessments
Chair: Frank Goldhammer, DIPF | Leibniz Institute for Research and Information in Education
Chair: Qiwei He, Educational Testing Service
Discussant: Alina A. Von Davier, ACT, Inc.

Computer-based assessments provide new insights into behavioral processes. A variety of timing and process data are recorded accompanying test performance data. Thus, much more data is available besides correctness or incorrectness. This symposium highlights the advanced psychometrics used in five studies to address questions on how timing data and action sequences are related to task performance and how to use such information to interpret test takers' achievements in large-scale assessments such as the Programme for International Student Assessment (PISA) and the Programme for International Assessment of Adult Competencies (PIAAC). The first paper investigates the effect of item-level time limits on individual differences in efficiency measures. The second paper introduces Bayesian covariance structure modeling to examine relationships between accuracy and process data variables. The third paper focuses on using action sequences to identify test takers' generalized behavioral patterns across items. The fourth paper provides evidence of relationships between background variables and sequential patterns in problem-solving tasks. The last paper explores how to incorporate response times into population modeling in PISA. These studies show the promise of modern psychometric techniques that can be utilized to exploit timing information and explore log file data, to improve proficiency estimation in large scale assessments.

How Do Item-Level Time Limits Affect Individual Differences in Efficiency Measures?
Frank Goldhammer, DIPF | Leibniz Institute for Research and Information in Education; Ulf Kroehne, German Institute for International Educational Research; Carolin Hahnel, German Institute for international Educational Research (DIPF) and Centre for International Student Assessment (ZIB); Paul De Boeck, The Ohio State University

Bayesian Covariance Structure Modeling of Response Accuracy and Response Times
Jean-Paul Fox, University of Twente

Using Process Data to Identify Generalized Patterns Across Problem-Solving Items
Qiwei He, Educational Testing Service; Francesca Borgonovi, OECD; Marco Paccagnella, OECD

Mapping Background Variables With Sequential Patterns in Problem-Solving Tasks
Dandan Liao, American Institutes for Research (AIR); Qiwei He, Educational Testing Service; Hong Jiao, University of Maryland-College Park

Incorporating Response Times Into Population Modeling in Large-Scale Assessments
Hyo Jeong Shin, Educational Testing Service; Kentaro Yamamoto, Educational Testing Service; Lale Khorramdel, ETS; Frederic Robin, ETS; Matthias Von Davier, National Board of Medical Examiners; Harrison Gamble, ETS; Wei Zhao, ETS

Saturday, April 6, 2019
2:15 – 3:45pm, British Columbia, Paper Session

New Insights in Test Assembly

Discussant: Kirk A. Becker, Pearson

Developing a Passage Difficulty Value Method to Assemble Passage-Based MST Using ATA

Ye Ma; John Denbleyker, Houghton Mifflin Harcourt

This study introduces a Passage Difficulty Values (PDV) method to solve issues involving the feasibility of mixed-integer linear programming (LP) approaches in ATA when assembling passage-based tests having large numbers of constraints. This methodology increases the possibility of using LP leading to the optimal assembly of tests.

Utility of Differential Functioning Statistics to Assess Item Position Effects

Mengyao Zhang, National Conference of Bar Examiners; Mark R. Connally, National Conference of Bar Examiners; Mark A. Albanese, National Conference of Bar Examiners

This study examines the utility of differential item and test functioning statistics to assess item position effects. Two commonly used differential functioning statistics are compared for data from a large-scale licensing test with multiple forms containing items in different orders. Both theoretical and practical implications of the findings are discussed.

Clustering Strategy for Assembling Abridged Test Forms From Retired Operational Test Forms

Yung-Chen Hsu, GED Testing Service; Tsung-Hsun Tsai, NBOME

Using retired operational forms to assemble multiple abridged practice forms is a common practice. Items are selected from a small pool, yet the forms still need to capture the essential test statistical characteristics. By employing a nearest neighbor clustering algorithm, we propose an effective approach to automate this pragmatic task.

A Linear Position-Discrepancy Response Time Model

Luping Niu, The University of Texas - Austin; Xiao Luo, Measured Progress; Louis A. Roussos, Measured Progress

The present study introduced a linear position-discrepancy response time model for detecting the linear effect of item position distance on the mean time spent when position shift exists. Results from the proposed model can inform test assembly of the acceptable position distance for negligible position effect on response time.

Automated Test Assembly for Multistage Testing With Cognitive Diagnosis

Guiyu Li, China Institute for Educational Finance Research, Peking University; Dongbo Tu, Jiangxi Normal University; Shaoyang Guo, East China Normal University

This study applies the computer multistage adaptive test (MST) to the cognitive diagnosis (CD) test (CD-MST) via Normalized Weighted Absolute Deviation Heuristic (NWADH) algorithms. The simulation indicates CD-MST has advantages of both MST and CD.

Saturday, April 6, 2019
2:15 – 3:45pm, Imperial Room, Electronic Board Session

Electronic Board Session 2

Performance and Response Behaviors on Multiple-Choice Questions: A Response Process-Based Validation Study

Michelle Chen, Paragon Testing Enterprises

Responding to the call for process-based approaches to test validation and extending on previous research on answer-changing behaviors, this study explores whether and how test-takers' response behaviors are associated with their performance. Log data of answer-selection behaviors from a listening comprehension test were analyzed using explanatory item response models.

The Comparability of Medical College Admission Test® Scores Obtained With Standard Versus Extra Time

Cynthia Anne Searcy, Association of American Medical Colleges; Marc Howard Kroopnick, Association of American Medical Colleges; Ying Jin, Association of American Medical Colleges

This study examined the comparability of scores from the Medical College Admission Test (MCAT) obtained under different amounts of administration time in relation to passing a medical licensing examination. These results suggest that some time accommodations may result in scores with equivalent meaning, and others may not.

Measurement Invariance in Classroom Climate Surveys

Meredith Langi, University of California - Los Angeles; Jonathan Schweig, The RAND Corporation; Jose Felipe Martinez, University of California - Los Angeles

To understand gender differences in classroom climate surveys, where data are multilevel and grouped at level-1, measurement invariance must fi be tested. The multilevel multiple group model (Asparouhov and Muthén, 2012) can test invariance in these data and measure group differences. We compare these results to those of fixed-effects models.

Construction of an Item Response Theory–Based Socioeconomic Status Index for NAEP 2003–2017

Youngjun Lee, Michigan State University; Yifan Bai, American Institutes for Research; Markus Broer, AIR

This study intends to construct a comparable SES scale for NAEP administrations 2003-2017. A multiple-group unidimensional IRT approach with mixed format items is adopted. Preliminary results suggest that the SES scale provides a comprehensive measure of students' family background and explains a significant amount of variance in students' achievement.

A Process for Using Content Alignment Indices in Instrument Revision

Anne Traynor, Purdue University; Tingxuan Li, Purdue University; Shuqi Zhou, Purdue University

We propose a two-step process for evaluating achievement test content alignment with curricular standards documents using judgmental item-objective matches assigned by subject-matter expert panelists. The method is demonstrated using item-objective match data from large-scale achievement tests in 13 US states.

Application of Wald Test to Detect Differential Item Functioning in Testlet-DINA Model

Wei Xu, National Council of State Boards of Nursing; M. David Miller, University of Florida

A growing number of studies have utilized CDMs to investigate DIF. Few studies, however, have adopted proper CDMs to detect DIF on testlet-based items. To address this gap, we use testlet -DINA model and implement Wald Test procedure to assess DIF items nested within a testlet.

Screening for Spiraling Irregularities in Random Groups Equating Design

Ying Lu, College Board; Judit Antal, The College Board; Sunhee Kim, College Board

Under random groups equating design, spiraling is commonly implemented to create randomly equivalent groups. When there are spiraling irregularities, however, the group equivalency assumption may be violated. This study investigates several purification procedures to identify spiraling irregularities in the equating sample and evaluates their impact on equating accuracy.

Careless Responding by Early and Late Semester Subject Pool Respondents

Theresa Trieu, Uniformed Services University of the Health Sciences; Javari Fairclough; Scott Plunkett, California State University - Northridge

This study examined whether late semester (last 3 weeks) subject pool participants different from early semester participants (first 3 weeks) on careless responding (i.e., miss more bogus items and attention check items). Effort devoted to the study, whether researchers should use the data, gender, and stress effects were also examined.

Fitting Models for Accuracy and Speed for Classification Into Learning Progression Levels

Peter Van Rijn, ETS Global; Edith Aurora Graf, ETS

In this study, item responses were mapped onto levels of a learning progression (LP) and item response theory models were used to evaluate these mappings. We also included response time in the modeling and found that this can lead to an improvement in the consistency of LP level classifications.

Exploration of Approaches Improving Item Parameters Estimations in Rapid Guessing Context

Rong Jin, Riverside Insights; John Denbleyker, Houghton Mifflin Harcourt; JP Kim, Riverside Insights

The low test-taker motivation is an issue in field test events. The resulted rapid guesses bring noises into item calibrations. This study will explore several calibration approaches on the robustness to rapid guesses by evaluating item discrimination and difficulty estimates and their standard errors.

Representing Quantitative Findings to Enhance Understanding: A Case From Elementary Science

Amy Cardace, UC - Berkeley; Kathleen E. Metz, University of California - Berkeley; Mark R. Wilson, University of California - Berkeley

This paper illustrates how visualizations can facilitate interpretation of measurement findings, especially focused on communicating with qualitative researchers about quantitative analyses. Our collaborative project in elementary science education presents a rigorous example of pre-post analysis while incorporating qualitative and quantitative perspectives to enhance learning and communication.

A Picture Is Worth a Thousand Words: An Investigation of Image-Based Assessments

Lisa Keller, University of Massachusetts; Jennifer Lee Lewis, University of Massachusetts - Amherst

Innovative assessments and Universal Design principles are hot topics in assessment today. One type of innovative assessment is an image-based assessment, that is free of text. This type of assessment eliminates language or literacy factors. This study is a preliminary investigation of basic design principles in constructing image-based assessments.

Understanding Unintended Consequences From Licensure Examination Score Interpretations: An Example From Medicine

Lauren Foster, National Board of Medical Examiners; Monica M. Cuddy, National Board of Medical Examiners

Consequences of licensure examination scores are important to validity arguments. Using logistic regression for a sample of 8,893 students/trainees, analyses examine the effect of secondary uses of medical licensure examination scores on physicians' career outcomes. Results suggest that for some students, scores may impact professional opportunities in unintended ways.

Learning Progression Validation Using Distractor-Driven, Multiple-Choice Items

Rajendra Chattergoon, University of Colorado - Boulder

This paper compares three methods for interpreting student response data from distractor-driven, multiple-choice items by describing the kinds of evidence each method could produce for the purpose of validating learning progressions. Preliminary results suggest that different scoring approaches and models yield contrasting information about the validity of a learning progression.

Patterns of Response Times for Different Student Strategies: Implications for Psychometric Modeling

Edith Aurora Graf, ETS; Peter van Rijn, ETS

A learning progression assumes that its levels are both ordered and distinct. These assumptions can be examined with IRT models, which can also be used for classification. Response times, and the strategies associated with them, are a largely untapped source of evidence that has the potential to improve level classifications.

Computerized Adaptive Testing With Hand-Scoring Items

Ching-Wei D Shin, Pearson; Ye Tong, Pearson

CAT programs such as SBAC use hand-scoring constructed responses (HSCR) items in CAT to assess deeper understanding of content skills. Results showed that adding HSCR items decreases adaptivity. Adding HSCR items adaptively within the test has less impact to test results than adding them at the end of the test.

Comparing the Consequences of Various Measurement Error Presentations in Test Score Reports

Dorien Hopster-den Otter, University of Twente; Elske Muilenburg, Universityof Twente; Saskia Wools, Cito; Bernard P. Veldkamp, Universiteit Twente; Theo Eggen, Cito

This paper investigated (1) the extent to which presentations of measurement error in score reports influence teachers' decisions and (2) teachers' preferences in relation to these presentations. Three presentation formats of measurement error (blur, colour value, and error bar) were compared to a presentation format that omitted measurement error.

Simulating Test Security Scenarios to Guide Application of DPF/DIF Forensics Method

James Davis, University of North Carolina - Greensboro; Andrew D Dallas, National Commission on Certification of Physician Assistants

This study illustrates the use of simulation in applying a forensics method to a particular testing context. The method involves DPF and DIF analyses to detect examinee preknowledge and item compromise (Smith & Davis-Becker, 2011). Results provide guidance on appropriateness of method and criterion choice given various security threat scenarios.

Evaluating Robustness of Reliability of Test Scores From Populations of Different Abilities

Lin Wang, ETS; Tsung-Han Ho, ETS

This study evaluated the IRT based reliability estimates from both pre-equated baseline tests and post-equated simulated tests from different population ability distributions. The pre-equated baseline tests and the post-equated simulated tests showed similar reliability estimates, suggesting the robustness of the reliability estimates from populations of different ability distributions.

Investigating the Impact of Item Pool Characteristics on Computerized Adaptive Testing

Yi He, ACT; Chunxin Wang, ACT, Inc.; Stephanie Su, ACT

This study investigates the impact of item pool characteristics, such as pool sizes, item difficulties, and content distribution, on the performance of a fixed-length computerized adaptive test (CAT). Results will provide information on the adequate pool characteristics that yield a CAT with desired measurement precision as well as item exposure.

A Three-Factor Model That Unifies Response, Time, and Missing

Ru Lu, Educational Testing Service; Hongwen Guo, ETS

This study proposes a three-factor model that expands the response time model to include missing into the measurement model. With an empirical data from a large-scale assessment, the proposed model is compared with three other models that do not account for missing data, to evaluate the added values.

Confirmatory Item Parameter Drift Detection in Computer Adaptive Testing

Kevin James Cappaert, Curriculum Associates; Yao Wen, Educational Records Bureau; Yu-Feng Chang, Minnesota Department of Education

Power and Type I error rates of confirmatory item parameter drift (IPD) detection methods are investigated. A quadrature point and weight adjusted D2 method and pseudo-count D2 and robust z methods are compared. All three have been found to result in acceptable Type I error rates and adequate power.

Practical Implications of Two Combined Subtests in Subscore Reporting in Multidimensional Item Response Theory

Yoon-Jeong Kang, AIR; Ming Li, Georgetown University

This study investigates practical implications of using subscores estimated from combined subtests for score reporting purposes. Results show that scores from combined subtests do not always represent students' true performance on each subtest, questioning the practice of combining subtests with small number of items to improve subscore reliability.

Comparisons of Text Complexity Across Groups of Summative Assessment Reading Passages

Tim Hazen, Iowa Testing Programs; Juliana Pacico, University of Massachusetts - Amherst

Utilizing a database of Reading passages at Grades 3-11, this study will examine the consistency of quantitative text complexity scores across groups of Reading tests. These results provide empirical evidence with which to test assumptions around the comparability of Reading rigor across a variety of assessment vehicles and audiences.

National Council on Measurement in Education Invited Electronic Board 2

Andre A. Rupp, Educational Testing Service (ETS)

Saturday, April 6, 2019
2:15 – 3:45pm, Manitoba, Coordinated Session

Assessment Literacy: What Do They Want to Learn? Four Perspectives

Chair: John Fremer, Caveon Test Security

In this session the presenters address the importance of assessment literacy and review strategies for teaching key players what they want to learn. The groups considered are: Parents, Teachers, School Administrators, Policy Makers. A passionate and long-term advocate for Assessment Literacy proposes an ambitious and workable framework. Direct experiences are shared from these perspectives: State Assessment Leadership; College level transition, research, and advocacy; The special case of situations where the possibility of cheating on tests has been raised. We will share what we have learned and how this could help others wanting to enhance Assessment Literacy in schools, colleges, state legislatures, and our overall society. Some key points in this session: Importance of Assessment Literacy, Similarity and differences of interests of groups, Need for collaboration to achieve assessment literacy, Strategies that can be employed and how they have worked, The use of brief messages. It is our intention to encourage contributions during our session by attendees regarding strategies they have tried, obstacles they have had to face and overcome, or anything else that is elicited by our presentations.

Assessment Literacy: What Do They Want to Learn — Advocate View
W. James Popham, University of California - Los Angeles

Assessment Literacy: What Do They Want to Learn — State View
Vincent M Verges, Florida Department of Education

Assessment Literacy: What Do They Want to Learn — Higher Ed Perspective
Michelle Croft, ACT, Inc.

Assessment Literacy: What Do They Want to Learn — Testing Industry Perspective
John Fremer, Caveon Test Security

Saturday, April 6, 2019
2:15 – 3:45pm, Quebec, Coordinated Session

Strengthening the Meaning and Utility of Test Scores for Their Intended Uses

Chair: Elizabeth Anne Summers, edCount, LLC
Discussant: Elizabeth Anne Summers, edCount, LLC

Addressing questions about the validity and reliability of assessment scores is an essential obligation of any person or agency using test scores to make judgments about any individual or group. This obligation applies whether a test is teacher-made for a class or produced commercially for large-scale use. Statewide and local assessments comprising an assessment system must be thoroughly evaluated for quality, standards and instructional alignment, purpose, utility, and equity, with an intentional focus on identifying and eliminating assessments that yield information that is (a) ambiguous or only interpretable in an ordinal (more than last time) sense; (b) simply overlapping or gained elsewhere; or (c) not connected to specific high-quality decisions and uses in combination with other data. Such efforts will strengthen the meaning and utility of test scores for their intended uses, and will consequently promote stakeholder understanding of the characteristics and benefi of a comprehensive assessment system. During this session, presenters will share a compilation of resources from the federally-funded SCILLSS project that are designed to help strengthen the knowledge base among state and local educators in the principles for high quality assessment that are critical to the appropriate selection, development, and use of assessments in educational settings.

Ensuring Rigor in State and Local Assessment Systems: A Self-Evaluation Protocol
Andrew Wiley, ACS Ventures, LLC

The SCILLSS Digital Workbook on Educational Assessment Design and Evaluation
Ellen E. Forte, edCount, LLC

The Benefits, Challenges, and Lessons Learned From Using the SCILLSS Resources
Rhonda True, Nebraska Department of Education; Charity Flores, Indiana Department of Education

A User's Perspective of Implementing the Local Self-Evaluation Protocol
Shannon Nepple, Adams-Central School District in Nebraska

Saturday, April 6, 2019
2:15 – 3:45pm, Salon A, Coordinated Session

Updating Career Readiness Assessments: Strategies, Challenges, and a Multimethod Validation Approach

Chair: Wayne J. Camara, ACT, Inc.
Discussant: Michael C. Rodriguez, University of Minnesota

The session highlights how ACT utilized Principled Assessment Design to update three assessments measuring career readiness. ACT defined career readiness as the attainment of work-related foundational cognitive skills. Since ACT developed their initial workplace assessments 25 years ago, significant changes have transformed the workplace and assessment design and validation. The goal was to develop updated assessments measuring today's foundational cognitive skills through the application of state-of-the-art design and validation practices. Validity considerations must be integrated in all design decisions. To illustrate, the session begins with an overview of the issues, challenges, and risks inherent to assessment revisions. The overview is followed by a review of ACT's application of the Principled Assessment Design approach utilized for the update. The session continues by detailing the multi-method validation approach whereby the team integrated five sources of validity evidence. The first report provides evidence collected through eye tracking studies that support the proposed cognitive processes used to solve assessment tasks. The second report provides psychometric evidence related to the assessments' internal structures, measurement precision, and fairness. It also provides ACT's continual plans to collect evidence analyzing the relationship of scores to critical outcome variables.

Maintaining the Long-Term Quality of Assessments in the Face of Change
Thanos Patelis, HumRRO

Principled Assessment Design: Applications and Tools for Assessment Updates
Thomas Langenfeld, ACT; Xiaohong Gao, ACT, Inc.

Using Eye-Tracking Data to Validate the Cognitive Processes of Foundational Workplace Skills
Jay Thomas, ACT, Inc.; Tom E. Langenfeld, ACT, Inc.

Integrating Multiple Sources of Psychometric Evidence to Support Assessment Update and Validation
Xiaohong Gao, ACT, Inc.; Chunyan Liu, National Board of Medical Examiners; Rongchun Zhu, ACT, Inc.; Meichu Fan, ACT, Inc.

Saturday, April 6, 2019
2:15 – 3:45pm, Salon B, Paper Session

Challenges in Standard-Setting

Discussant: Daniel Lewis, ACT

Evaluating Panelists' Understanding of Standard-Setting Data

Patricia Baron, ETS; Sharon Cadman Slater, ETS; Stephen G. Sireci, University of Massachusetts - Amherst

Survey responses from eight panels of teachers participating in a Bookmark standard setting showed that data presented during the process was misunderstood by some panelist, and tended to be misused. Results provide some themes around misconceptions observed and the need for training and evaluation around data used in standard setting.

When Consequences of False Negative Misclassification Have Greater Harm: A 3-PL Illustration

Brian Leventhal, James Madison University

The Weighted Classification Error (WCE) function quantifies misclassification error rates for assessments scored using the three-parameter logistic model. Applied to a high-stakes information literacy examination, the WCE quantifies the weighted likelihood of harm when a false negative error is considered more harmful than a false positive error.

The Choice of Response Probability in Bookmark Standard-Setting: An Experimental Study

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

Cut scores bookmark standard setting exercises should reflect judges' opinions about the trait level needed for a given classification and not the response probability they are instructed to use. Results from randomly assigning judges to one of two response probability conditions showed systematic cut-score differences across groups.

Assessing at the Very Beginning (a Very Good Place to Start)

Leslie Keng, Center for Assessment; Joseph A. Martineau, Center for Assessment; Cydnee Carter, Utah State Board of Education; Jennifer Throndsen, Utah State Board of Education

This paper details the efforts that one state undertook to design and implement assessments intended to measure its kindergarten students in early literacy and numeracy. The session will cover four key aspects of the new assessments: item and test development, calibration and scaling, standard setting, and reporting.

Saturday, April 6, 2019
2:15 – 3:45pm, Territories, Paper Session

The Role of Student Interest and Engagement on Performance

Discussant: Dianne Henderson, Renaissance Learning, Inc.

Investigating the Strategic Allocation of Time on Task in a Computer-Based Assessment

Johannes Naumann, University of Wuppertal

Interactive effects of each print reading skills, strategy knowledge, and reading enjoyment with task difficulty on time on task-relevant pages and total time on task in digital reading were estimated using PISA 2009 data. Results suggest that skilled, knowledgeable and motivated students acted more adaptively in allocating time on task.

Examinee Test-Taking Effort on an Adult Proficiency Test

Sandra Botha, University of Massachusetts - Amherst

Low examinee effort presents a validity threat to inferences made from test scores. This study used response time data from an adult proficiency test to investigate examinee engagement and explored the use of item response times as a method of identifying test items that consistently receive low effort from examinees.

Diverse Engagement Profiles: Demonstration and Implications of Test Preparation for High-Stakes Exams

Abeer A. Alamri, University of South Florida; Edgar I. Sanchez, ACT, Inc.

Latent profi analysis was used to characterize usage of a test preparation program. Three profi emerged (low-usage/ low-performance, high-usage/ moderate-performance, and low-usage/ high-performance) with overrepresentation of females in low-usage/low-performance and high-usage/moderate-performance profiles. Findings inform our understanding of student engagement with test preparation and can inform communication efforts at improving engagement.

Academic Mind-Sets, Engagement, and Academic Performance of Fourth Graders' Reading, Mathematics, and Science

Ze Wang, University of Missouri

Using 4th-grade U.S. samples from TIMSS 2011 and PIRLS 2011, this study tested a hypothesized model that Academic Mindsets influence academic performance through Academic Behaviors in reading, mathematics, and science. Results suggest that the theoretical model was in general supported at the student level but not at the classroom level.

Saturday, April 6, 2019
4:10 – 6:10pm, Alberta, Invited Speaker Session

NCME Session on Excellence in Public Communications
Panelist: Amy Burkhardt, University of Colorado - Boulder
Panelist: Ellen E. Forte, edCount, LLC
Panelist: Min Li, University of Washington
Panelist: W. James Popham, University of California - Los Angeles
Panelist: Javarro Antoine Russell, Educational Testing Service
Panelist: William Skorupski, Amira Learning
Panelist: Catherine Gewertz, Education Week

Each of us has been frustrated by inaccurate measurement information in the media and by faulty policies based on a lack of measurement acumen. To help improve the quality of published information about testing and, thus, improve testing practices and uses of test scores, the NCME Board established a new awards committee to recognize, honor, and encourage excellence in the public communication of measurement information to stakeholders outside of our measurement field. The first award will be announced during the 2019 annual meeting and the award winner will participate in this special session. This session will offer insights into the award purpose and the nomination and review criteria as well as examples of excellence in public communications. Such communications, as recognized in the criteria for this new award, will include particular works or collections that explain educational measurement concepts and their signifi e to groups of education stakeholders (e.g., administrators, teachers, parents) and others (e.g., policy-makers, business leaders, the general public). Throughout the discussions of these works, we will highlight characteristics of excellence in public communications and reflect upon how all members of our measurement community can contribute to improving the quality of our public-facing professional discourse.

Saturday, April 6, 2019
4:10 – 6:10pm, Algonquin, Coordinated Session

Utilizing Expert Judgments to Facilitate Scaling of Tests Adapted for Small Populations

Chair: Mark Hansen, University of California - Los Angeles
Discussant: Stephen G. Sireci, University of Massachusetts - Amherst

Making changes to or sometimes replacing test items is often necessary to address student accessibility needs and to ensure fair testing conditions. With such changes, however, scoring parameters from one assessment format can be of questionable relevance to the alternative format. Furthermore, the number of students taking an alternative form of an assessment is sometimes far too small to conduct a traditional item-based linking study. It is nonetheless critical that performance on alternative assessment forms support the same inferences about student performance. Within this session, we explore methods of scaling alternative test forms that have been developed for small populations. Our motivating example is the Braille version of a K-12 English Language Proficiency (ELP) assessment. We begin by discussing the development of this test, then present alternative approaches utilizing expert judgments to scale the assessment. The proposed methods are examined and contrasted with current practices related to the scoring of the Braille versions of ELP tests. Finally, we propose some general principles for the design of data collections that support these novel approaches to item calibration.

On the Challenges of Scaling Tests for Small Populations
Kurt T. Taube, Ohio Department of Education; Mark Hansen, University of California - Los Angeles

Using Expert Judgments to Estimate Scoring Parameters
Mark Hansen, University of California - Los Angeles; Phoebe C. Winter, Self-employed; Michelle McCoy, University of California - Los Angeles; Nami Shin, University of California - Los Angeles

Moderated Item Calibration
Seungwon Chung, University of California - Los Angeles; Li Cai, University of California, Los Angeles

Optimal Design in Rating Procedures Involving Test Items
Li Cai, University of California, Los Angeles; Sijia Huang, University of California - Los Angeles

Saturday, April 6, 2019
4:10 – 6:10pm, Ballroom, Coordinated Session

Automated Scoring Validity Research for a National Large-Scale Writing Assessment

Chair: Scott William Wood, ACT, Inc.
Discussant: Susan Marie Lottridge, American Institutes for Research

Adding automated scoring to an established writing assessment can provide many benefi to stakeholders including faster scoring times and reduced scoring costs. At the same time, the use of automated scoring on an established writing assessment raises many validity concerns from its stakeholders. Stakeholders expect that sufficient research is conducted to determine the validity of automated scoring's use. This session provides an overview of automated scoring research demonstrating the validity of using the CRASE engine to score Australia's National Assessment Program – Literacy and Numeracy (NAPLAN) Writing Assessment, specifi addressing concerns regarding the strength and structure of validity evidence raised by its stakeholders and the public. Presenters will demonstrate four areas of automated scoring research conducted using pilot and operational data that address specific gaps in validity evidence structure and gather new empirical evidence to engage constructively and productively with the raised concerns, increasing chances of the future use of automated scoring in the NAPLAN Writing Assessment. The research findings presented in each presentation serve as an example for other writing assessments of how to respond to and communicate concerns raised by stakeholders regarding the validity of using automated scoring.

National Assessment Program–Literacy and Numeracy (NAPLAN) Writing Automated Scoring and Stakeholder Validity Concerns
Goran Lazendic, Australian Curriculum, Assessment and Reporting Authority

Identifying Nonvalid Examinee Scripts Automatically
Alejandro Andrade, ACT, Inc.; Gavin Henderson, ACT, Inc.; Erin Yao, ACT, Inc.

Methods for Detecting When Examinees Game an Automated Scoring Engine
Gavin Henderson, ACT, Inc.; Alejandro Andrade, ACT, Inc.

Ensuring Automated Scoring Fairness for Key Subgroups
Scott William Wood, ACT, Inc.

Reducing Automated Scoring Training Costs Through Generalized and Pseudo-Generalized Models
Erin Yao, ACT; Scott William Wood, ACT, Inc.

Saturday, April 6, 2019
4:10 – 6:10pm, British Columbia, Paper Session

The Need for Speed? Practical Assessment Implications

Discussant: Wayne J. Camara, ACT, Inc.

Modeling Speededness in Medical Licensure Examination Administered Under Varying Timing Constraints

Chunyan Liu, National Board of Medical Examiners; Wenli Ouyang, National Board of Medical Examiners; Polina Harik, National Board of Medical Examiners

Test speededness is an important consideration in test design. The purpose of this study is to determine the optimal testing time for a medical licensure examination using lognormal response time modeling. In addition, we will investigate the relationship between response time and examinee ability, gender and native language.

Patterns of Pacing Behavior in a Medical Licensing Examination

Wenli Ouyang, National Board of Medical Examiners; Matthias Von Davier, National Board of Medical Examiners; Polina Harik, National Board of Medical Examiners

Initial results of a study aiming at identifying patterns of pacing behavior using a latent class mixed modeling approach are presented. The results were validated by comparing class specific distributions of examinee characteristics such as repeater status, medical school accreditation, native language and gender.

Can the Effect of Test Speededness Be Mitigated Through Item Purification?

Yage Guo, University of Nebraska - Lincoln; Richard A. Feinberg, National Board of Medical Examiners; Chunyan Liu, National Board of Medical Examiners

Undesired speededness presents an immediate problem when it is detected on the current administration of a test. This study proposes a new method to lessen the effects of speededness by item purification that shows promise in recovering ability parameters through a comparison of experimental timing conditions.

Exploring Perceived Time Pressure and Speededness Using NAEP Process Data

Fusun Sahin, American Institues for Research; Juanita Hicks, American Institutes for Research; Mingqin Zhang, The University of Iowa; Cheng Shuang (Grace) Ji, American Institutes for Research

This study uses process data to validate examinees' self-reported time-pressure in the 2017 NAEP Mathematics Assessment. Examinees who perceived "a lot of" time pressure did indeed show increasingly speeded behavior and interacted less with items towards the end of the test, providing an objective validation of this contextual questionnaire item.

Two Methods for Detecting Rushing Behavior in Adaptive Testing

Can Shao, Curriculum Associates; Logan Andrew Rome, Curriculum Associates

This paper uses two methods to detect rushing behavior in a Computerized Adaptive Test (CAT): change-point analysis and simple descriptive statistics. The first method will focus on examinees' response patterns while the latter will consider response time and accuracy. A detailed comparison of these two methods will be discussed.

Saturday, April 6, 2019
4:10 – 6:10pm, Manitoba, Coordinated Session

Beyond Learning Progressions: Maps as Assessment Architecture

Chair: Meagan Karvonen, The University of Kansas
Discussant: James W. Pellegrino, University of Illinois at Chicago

Learning progressions (LPs) are commonly used in educational assessments to identify interim steps on a pathway toward a grade-level target. LPs describe typical expected pathways, but may not represent the multiple pathways by which students develop knowledge in a domain. Another type of cognitive model, the learning map, is better suited to describing heterogeneous pathways that support learning for all students including those with the most significant cognitive disabilities. This session ties together four presentations on different facets of a project involving the creation and use of maps as cognitive learning models to support the design of large-scale assessments. The first presentation illustrates how an assessment's theory of action and validity argument are grounded in the maps as models of the content domains. The second presentation describes the map creation process, including intentional design decisions and the application of universal design for learning principles. The third presentation describes the iterative design process and the use of stakeholder evaluation processes to evaluate the maps for content and accessibility. The fourth presentation describes empirical methods for map validation. The session ends with discussion of lessons learned and future directions, and commentary from a national expert in cognitive learning models and large-scale assessment.

Grounding the Design of a Large-Scale Alternate Assessment in Learning Map Models
Meagan Karvonen, The University of Kansas; Russell E. Swinburne Romine, The University of Kansas

Learning Maps as Models of the Content Domain
Russell E. Swinburne Romine, The University of Kansas; Jonathan Schuster, The University of Kansas

Iterative Design and Stakeholder Evaluation of Learning Map Models
Lori Andersen, University of Kansas; Russell E. Swinburne Romine, The University of Kansas

Empirical Methods for Evaluating Maps: Illustrations and Results
Jake Thompson, The University of Kansas; Brooke Nash, The University of Kansas

Saturday, April 6, 2019
4:10 – 6:10pm, Salon A, Paper Session

Communicating Performance Results to Various Audiences

Discussant: Rochelle S. Michel, Educational Records Bureau

An Argument for Reporting the Likelihood Function in Large-Scale Survey Assessments

John R Lockwood, Educational Testing Service

Large-scale survey assessments such as NAEP, PISA and PIAAC report ``plausible values''of respondents'latent traits intended for secondary analysis. We will argue why the likelihood function of latent traits given item responses also should be reported. Benefits will be demonstrated using NAEP data for small-area estimation.

Communicating Process Data to Teachers for Conversation-Based Assessment

Stephanie Peters, Educational Testing Service; Dr. Carol McGregor Forsyth, Educational Testing Service; Diego Zapata-Rivera, Educational Testing Service

Creating score reports is a complicated process requiring iterative refinement particularly for simulations producing process data. To create score reports with process data, we consulted experts in multiple fields of research. Communicating these reports to teachers requires iterative input from teachers to discover the best measures to present.

Indices of Conditional Measurement Precision and the Impact of Scale Transformation

Dongmei Li, ACT, Inc.

Conditional standard error of measurement (CSEM) is widely used in educational measurement. Prompted by recent challenges against the value and adequacy of reporting CSEM, this study demonstrates the impact of scale transformation on CSEM to help clarify expectations and misconceptions in CSEM calculation and interpretation.

Using Finite Mixture Models to Communicate Results Across Multiple Assessments

Margarita Olivera Aguilar, Educational Testing Service; Samuel Rikoon, Educational Testing Service

We propose and illustrate using fi e mixture model results to design score reports aiming to communicate scores across multiple scales. We emphasize the simplification of complex information via a visual display. We also compare feedback about the clarity and usefulness of such score reports vs. traditional reports.

Saturday, April 6, 2019
4:10 – 6:10pm, Salon B, Paper Session

Equating: Applications and Insights

Discussant: Won-Chan Lee, University of Iowa

An Investigation of Observed Score Anchor Construction Practices in Certification/Licensure Testing

Joshua David MacInnes, Scantron Corporation; Richard M. Luecht, University of North Carolina - Greensboro; Devdass Sunnassee, University of North Carolina - Greensboro; Randall D. Penfield, University of North Carolina - Greensboro; John T. Willse, University of North Carolina at Greensboro

This simulation study examined anchor construction practices in observed score equating under certification/licensure testing conditions. The results identified important item difficulty and examinee ability interactions that impact error, particularly when one sample of examinees is more homogeneous. A set of guidelines for practitioners were developed based on the results.

Impact of Rasch Item Parameter Drift in Small Samples Over Multiple Administrations

Jason P. Kopp, American Board of Surgery; Andrew Jones, American Board of Surgery

The current study simulated data using the Rasch model to investigate the impact of item drift with small sample sizes of 25 and 50. Classification accuracy was strongly degraded by item drift of 0.5 logits or greater, suggesting unaddressed item drift can substantially affect pass-fail decisions under small sample conditions.

Linking to a Calibrated Item Pool With Short External Anchor Tests

Seohee Park; Michael E. Walker, The College Board; Sunhee Kim, College Board

This study focuses on an equating design where the equating items are administered in a matrix fashion such that any individual examinee receives only a part of the external anchor test; but where linking is implemented with the entire anchor test by aggregating over examinees.

Inclusion of Constructed-Response Items in Anchor Sets

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

As assessments expand beyond traditional multiple-choice tests, consideration should also be given to the expansion of anchor sets for equating. The purpose of this study is to explore the impact of the inclusion of constructed-response items in the anchor set on equating results.

Rasch Versus Classical Equating in the Context of Small Sample Sizes

Ben Babcock, The American Registry of Radiologic Technologists; Kari Hodge, NACE International Institute

We extend past small-sample equating research by 1) directly comparing classical and Rasch techniques and 2) pooling multiple forms' worth of data to improve Rasch estimation. Results showed that combining multiple administrations' data via the Rasch model yields more accurate equating compared to small-sample classical methods.

Saturday, April 6, 2019
4:10 – 6:10pm, Territories, Invited Speaker Session

Assessment Literacy: Tactics for Traction and Strategies for Success

Panelist: Michael D. Beck, Beta, Inc.
Panelist: Christopher R. Gareis, College of William and Mary
Panelist: Thomas R. Guskey, University of Kentucky
Panelist: Susan B. Nolen, University of Washington - Seattle
Panelist: W. James Popham, University of California - Los Angeles
Panelist: Kecia L. Addison, Montgomery County Public Schools
Moderator: Stephen C Court, CRESST

Assessment literacy has become more prominent and pressing this decade. The topic is discussed more widely and more frequently. Articles appear in not only academic journals but also popular periodicals and online postings. Assessment literacy surveys and self-assessments circulate. Standards and performance measures have been formulated. ESSA funds can be leveraged to support assessment literacy initiatives at the state and local levels. Task forces and ad hoc alliances have been formed. Entire websites are now devoted to the topic. Some include sets of online instructional modules and other useful resources intended to raise the assessment literacy levels of their viewers. NCME is rating the quality of these resources and will link acceptable ones to its website. Yet, despite the increase in salience, funding, and resources, assessment literacy levels remain woefully low among too many members of major stakeholder groups – from students and teachers to the U.S. Secretary of Education. A more concerted and effective effort is needed. As a national organization, the National Association of Assessment Directors (NAAD) seems perfectly positioned to coordinate an inter-organizational campaign to raise assessment literacy levels. Accordingly, during this year's NCME/NAAD symposium, a panel of noted assessment literacy experts and district-level practitioners will discuss goals, strategy, tactics, and timelines for such a campaign. Audience input will be welcomed.

Saturday, April 6, 2019
6:30 – 8:00pm, Concert Hall, Convention Floor

NCME and Division D Reception

Sunday, April 7, 2019
8:00am-10:00am, Concert Hall, Convention Floor

NCME Breakfast, Business Meeting, and Presidential Address

NCME Presidential Address

Fairness In Measurement and Selection: Statistical, Philosophical, and Public Perspectives

Rebecca Zwick
Educational Testing Service

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

Sunday, April 7, 2019
10:20 – 11:50am, Alberta, Coordinated Session

Evaluating Test Speededness in NAEP Digitally Based Assessments

Chair: Fusun Sahin, American Institutes for Research
Discussant: Can Shao, Curriculum Associates

Test speededness occurs when time constraints influence test takers' performance, which negatively affects the reliability and validity of large-scale power tests such as NAEP. To date, one popular rule for assessing speededness is the ETS's (1974) rule where"test is described to be unspeeded if at least 80% of test takers reached last item and if everyone reached at least 75% of the items."Yet, this common approach is criticized to be arbitrary and based on non-response rate only. With the rise of digitally based assessments, especially recordings of response times (RT) for each item have resulted in an increasing number of research studies investigating the speededness. Recently, change point analyses (CPA) based either on RT or response pattern have been proposed. This symposium presents and compares the results of three major speededness detection methods as applied to 2017 NAEP grade 4 and 8 mathematics DBA data. The methods include: CPA based on response time (Shao, 2016), CPA based on response patterns (Shao, 2016; Shao et al., 2016), and CPA based on cumulative sum scores (Sinharay, 2017). The results of the three methods will be compared and discussed. The characteristics of the identified speeded examinees will be examined.

Conceptual Framework and Overview of Speededness in NAEP
Young Yee Kim, American Institutes for Research

Detecting Speededness Using Change Point Analysis on Response Data
Xinyu Ni, Teachers College, Columbia University; Glenn Hui, George Mason University; Xiaying Zheng, American Institutes for Research

Applying CUSUM-Based Person Fit Statistics to Detect Speededness
Ummugul Bezirhan, Teachers College, Columbia University; Xiaying Zheng, American Institutes for Research

Detecting Speededness Using Change Point Analysis on Response Time
Mingqin Zhang, The University of Iowa; Xiaying Zheng, American Institutes for Research; Glenn Hui, George Mason University

Comparison of Results and Discussion
Xiaying Zheng, American Institutes for Research

Sunday, April 7, 2019
10:20–11:50am, Algonquin, Coordinated Session

New Challenges in Variance Estimation for Digital-Based Educational Assessment

Chair: John Mazzeo, ETS
Discussant: Derek C Briggs, University of Colorado - Boulder

It is critical to incorporate and estimate accurately all sources of error variances for the reported results of large-scale educational survey assessments, like NAEP, PISA, and TIMMS. Since the survey design of the assessments are typically complex, it can be hard for public users to calculate the statistical variances on their own. Therefore, getting the accurate variance estimations for the public users is an important part of communicating the results clearly and accurately to the public. New challenges arise when educational assessments are moving to digital testing platform, particularly when results from the digital platform are compared to previous paper results. Bridge studies are usually conducted to allow these programs to report results obtained from the digital assessments on the same scales as earlier paper results. Depending on the bridge design, extra variances due to linking error or item sampling should be included in the total variance. In this session, researchers working on National Assessment of Educational Progress (NAEP) will present papers focused on error variance estimation during NAEP transition from paper-pencil assessment to digital-based assessment. Although the research is conducted to solve challenges in NAEP, the presented are general to other large-scale educational assessment with similar design.

Error Variance in Common-Population Linking Designs With a Linear Transformation Function
Paul A Jewsbury, Educational Testing Service

New Jackknife Variance Estimators for NAEP Transitioning to Digital-Based Assessment
Bingchen Liu, Educational Testing Service; John Mazzeo, ETS

Variance Components in a Writing Bridge Study
Xueli Xu, Educational Testing Service; Yue Jia, Educational Testing Service

Issues in Finite Population Correction in Variance Estimation for Sampling Without Replacement
Jiahe Qian, ETS

Sunday, April 7, 2019
10:20 – 11:50am, Ballroom, Coordinated Session

Communicating and Reporting Student Growth

Chair: Katherine Furgol Castellano, Educational Testing Service
Discussant: Daniel F. McCaffrey, ETS

A fundamental component of many states' assessment and accountability systems include a student growth measure. This symposium offers insights into the challenges different states, testing companies, and researchers face in communicating and reporting student growth for uses at the student, teacher, school, and/or district level to key stakeholders from students and parents to teachers and administrators. Two states and a testing company reflect on these challenges through their histories of reporting student growth for different uses. A comprehensive review is also presented of tools developed across several states for communicating student growth to stakeholders, including an assessment of the extent that the effectiveness of these tools has been documented. In a similar vein, we present a method for visualizing measurement error in student growth measures and results for a small survey study of its utility and interpretability by teachers and administrators. The symposium also discusses considerations for different uses of student growth at different levels and reflects on gaps between state needs and current research in student growth. Ample time is included for open discussion between the presenters and symposium participants to further the discussion on best practice in reporting and using student growth.

The Evolving Story of Student Growth in California
Kimberly Mundhenk, California Department of Education

Communicating Student Growth Information in Georgia
Allison Timberlake, Georgia Department of Education

Lessons Learned From Ongoing Outreach in Communicating Student Growth
Eric Stickney, Renaissance Learning

A Review of How States Are Communicating Student Growth to Parents
Sharon Cadman Slater, ETS

An Effort to Communicate Measurement Error in Student Growth Reports
Katherine Furgol Castellano, Educational Testing Service; Daniel F. McCaffrey, ETS

Sunday, April 7, 2019
10:20 – 11:50am, British Columbia, Coordinated Session

Measurement and Communication Challenges in a Technology-Based Book Reading Intervention

Chair: Beata Beigman Kelbanov, Educational Testing Service
Discussant: Arthur Graesser, University of Memphis, Tennessee

Thirty-two percent of U.S. 4th graders read below the Basic level (NAEP, 2017), most of these children showing evidence of low reading fl y. With the advent of e-reading technology, new opportunities for integrating measurement with intervention in the development of reading skills are available. Specifically, we can measure e-book reading activity continuously, and use feedback to monitor progress and help mediate student engagement and interactions during reading of high-interest books. In this session, we report on results from pilots of MyTurnTo-Read – an e-book-based tool designed to support an interleaved listening and reading experience, where the child takes turns reading aloud with a virtual partner (audiobook). Reading data was collected from students in summer camps reading Harry Potter and the Sorcerer's Stone. The presentations address measurement and communication challenges that arise from (a) using a high-interest, not-grade-level-controlled book in a noisy setting; (b) relating observed reading patterns to reading skills; (c) engaging the readers with the system; (d) communicating measurements provided by the tool to teachers. Collectively, the research suggests that such technologically-situated book reading is promising as an activity that supports both sustained reading for meaning and pleasure and effective measurement of reading fluency and basic comprehension.

Measuring the Effect of Textual Features on Children's Oral Reading Performance
Van Rynald Liceralde, Educational Testing Service; Beata Beigman Kelbanov, Educational Testing Service; Anastassia Loukina, Educational Testing Service; John R Lockwood, Educational Testing Service

The Impact of Ambient Noise on Measurement of Oral Reading Performance
Anastassia Loukina, Educational Testing Service; Patrick Lange, Educational Testing Service; Qian Yao, Educational Testing Service; Beata Beigman Kelbanov, Educational Testing Service; Nitin Madnani, Educational Testing Service; Abhinav Misra, Educational Testing Service; Klaus Zechner, Educational Testing Service

Relating Performance in Extended Book Reading to Measures of Reading Skills
John P. Sabatini, ETS; Zuowei Wang, Educational Testing Service; Tenaha P. O'Reilly, ETS

The Impact of Task and Student Characteristics on Engagement During Collaborative Reading
Blair Lehman, Educational Testing Service

Evaluating Teachers' Needs for Ongoing Feedback From a Technology-Based Book Reading Intervention
Priya Kannan, Educational Testing Service; Beata Beigman Kelbanov, Educational Testing Service; Vera Shao, Educational Testing Service; Colleen Appel, Educational Testing Service; Rodolfo Long, Educational Testing Service

Sunday, April 7, 2019
10:20 – 11:50am, Quebec, Coordinated Session

Testing, Testing: Retesting and Inequality in Large-Scale College Admissions Tests (Diversity Issues and Testing Committee's Selected Session)

Chair: Andrew Ho, Harvard University
Discussant: Darryl Hill, Assistant Superintendent for School Accountability, Fulton County School District

As federal policies emphasize college access (Every Student Succeeds Act, 2015), the cost of college admissions test scores remains a barrier to equitable college access for low-income students (Dynarski, 2018). In a past NCME symposium on retesting (Bertling & Ho, 2017; Mattern & Radunzel, 2017), we described how college scoring policies create incentives for prospective students to retake exams. This symposium extends past work to quantify socioeconomic gaps in retest rates, estimate the causal impact of retesting on college admissions, and evaluate interventions designed to increase access to retesting for low-income students. Together, these papers ask whether and how admissions retesting policies and interventions can improve college access and close socioeconomic gaps. Paper 1 overviews trends and socioeconomic gaps in ACT retesting rates. Paper 2 evaluates the impact of statewide ACT adoption on retesting behavior, particularly for low-income students. Paper 3 uses a regression discontinuity design to estimate the causal impact of retaking the SAT on college enrollment. Paper 4 concludes with an experimental evaluation of an ACT fee-waiver policy for economically disadvantaged students. Darryl Hill will discuss these papers from his experience leading accountability and research efforts in Wake County, North Carolina, and Fulton County, Georgia.

Demographics and Differential Advantages in College Admissions Retesting
Maria Bertling, Harvard University; Andrew Ho, Harvard University

If the State Pays, Will Students Retest on Another Day? Impact of Statewide-Adoption of the ACT on Testing and Retesting Patterns
Krista D. Mattern, ACT, Inc.; Justine Radunzel, ACT, Inc.

Take Two! SAT Retaking and Inequality in College Enrollment
Joshua S. Goodman, Harvard University; Jonathan Smith, Georgia State University; Oded Gurantz, College Board

Three Experiments to Improve the College Entrance Exam Attendance Rates of Low-Income Students
Ty M. Cruce, ACT, Inc.; Robert Hahn, University of Oxford - Metcalfe; Robert Metcalfe, Boston University

Sunday, April 7, 2019
10:20 – 11:50am, Salon A, Paper Session

Applications of Social and Emotional Learning Measures

Discussant: Matthew Newman Gaertner, WestEd

Development and Scoring of a New Interpersonal Skills Situational Judgment Test

Samuel Rikoon, Educational Testing Service; Lisa Merrill, Research Alliance for New York City Schools

To improve communication of social emotional assessment results to students and stakeholders, a situational judgment test targeting Interpersonal Skills was developed for middle and high school students. We examine its psychometric properties in terms of scoring, reliability, internal structure, and validity evidence. Implications and future research plans are discussed.

Multilevel Reporting on Collaborative Problem Solving in an Educational Simulation

Dr. Carol McGregor Forsyth, Educational Testing Service; Stephanie Peters, Educational Testing Service; Jessica Andrews Todd, Educational Testing Service; Andre A. Rupp, Educational Testing Service (ETS)

Score reports can become increasingly complex with interactive environments containing multiple levels. In the current study, we investigated collaborative problem solving skills that best predict success on each level in an online electronics environment. These findings are an initial step in determining types of score reports for a multi-level simulation.

Multimethod Approach to the Measurement of Socio-Emotional Skills: A Unified Scoring Method

Cristina Anguiano-Carrasco, ACT, Inc.; Carrie Morris, ACT; Kate Walton, ACT, Inc.

The importance of socio-emotional skills is well known, but measuring such skills remains challenging. Using three different item types to measure Grit, we developed a psychometric model for generating a unified score. Correlations with GPA were higher with the unified score than with scores from each item type.

Does Valuing Collaboration Lead to Greater Success in Collaborative Problem Solving?

Chang Lu, University of Alberta; Okan Bulut, University of Alberta

This study examined attitudinal differences in collaboration between students from China and Canada in PISA 2015. Results from the explanatory, polytomous item response models indicated that Chinese students are more likely to collaborate than Canadian students, although Canadian students scored significantly higher than Chinese students in the collaborative problem-solving test.

Social-Emotional Learning ICCs and Associations With School Composition and Achievement

Kyle Nickodem, University of Minnesota; Michael C. Rodriguez, University of Minnesota; Rik Lamm, University of Minnesota - Twin Cities; Kyungin Park, University of Minnesota - Twin Cities

To explore the psychometric appropriateness for including measures of social and emotional learning as indicators of school quality, we estimate ICCs and the extent to which such measures are associated with school composition and school-level achievement. We find little support to endorse such use.

Sunday, April 7, 2019
10:20 – 11:50am, Salon B, Paper Session

Technical Considerations in Factor Analysis and Structural Equation Models

Discussant: Sarah Quesen, Pearson Education, Inc.

An Evaluation of Hierarchical Models Relating Item Response Format, Accuracy, and Speed

Xin Qiao, University of Maryland - College Park; Usama S. Ali, Educational Testing Service; Peter van Rijn, ETS

The current study evaluates innovative item types for computer-based tests using models for speed and accuracy in a confirmatory factor analysis/structural equation modeling framework using empirical assessment data. The results indicate that item features such as response format are related to item parameters for both accuracy and speed.

Negatively Worded Items in a Self-Report Multidimensional Measure

Feifei Ye, The RAND Corporation; Xiaoyan Xia, University of Pittsburgh

This study investigates the functioning of negatively worded items (NWI) in balanced scales by comparing a single bifactor model (NWI as a method factor) and a double bifactor approach (NWI as a substantive factor). Using empirical and simulated data, we examine the consequence of misspecifying the model for NWI.

Further Exploration of Vertical Scaling Using a Bifactor Item Response Theory Model

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

In vertical scaling, construct shift threatens the validity of inferences when the underlying scale is assumed unidimensional. To address this issue, Li and Lissitz (2012) proposed a multigroup bifactor model where grade-specific latent dimensions are modeled. This study furthers this line of research by considering a more general bifactor model.

Within-Item Interactions in Bifactor Models for Ordered-Categorical Item Responses

Meghan Fager, National University; Jonathan Templin, University of Kansas

This research introduces an extension of bifactor item response models to include an interaction effect at the item-level between general and domain-specific dimensions. Empirical and simulated data are studied to validate and test the proposed model and evaluate the potential adverse effects that may arise from omitting interactions.

Sampling Distributions of AIC and BIC Differences for Higher Order and Bifactor Models

William Skorupski, Amira Learning; Matthew Reynolds, University of Kansas

Various higher-order and bifactor models were simulated to evaluate the performance of AIC and BIC difference statistics for identifying the correct model. Simulations were varied to reflect differences in the true model, number of factors, indicators, and sample size for estimation. Heuristics to evaluate meaningful differences are offered.

Sunday, April 7, 2019
10:20 – 11:50am, Territories, Paper Session

New Directions in Cognitive Diagnostic Modeling

Discussant: Hong Jiao, University of Maryland-College Park

Cognitive Diagnostic Computerized Adaptive Testing (CD-CAT) for Small Educational Programs: A General Nonparametric Item Selection Method

Yuan-Pei Chang, Rutgers University - New Brunswick/Piscataway; Chia-Yi Chiu, Rutgers University

A general nonparametric item selection (GNPS) method for CD-CAT is proposed in the study. The algorithm can be used for items conforming to the saturated general CDMs and the result of the preliminary simulation shows that it outperforms the compared parametric methods when the calibration samples are small.

A New General and Effective Method of Q-Matrix Validation

Daxun Wang, Jiangxi Normal University; Wenchao Ma, The University of Alabama - Tuscaloosa; Xuliang Gao, Jiangxi Normal University; Yan Cai, Jiangxi Normal University; Dongbo Tu, Jiangxi Normal University

This study proposed a general method based on likelihood ratio test (LRT) to validate Q-matrix, which can be used with a wide class of cognitive diagnosis models (CDMs). Results showed that the proposed method on the whole outperforms the existing methods whatever the reduced or saturated CDMs are used.

Data-Driven Q-Matrix Validation Using a Residual-Based Statistic in Cognitive Diagnostic

Xiaofeng Yu, University of Notre Dame; Ying Cheng, University of Notre Dame; Alex Brodersen, University of Notre Dame

This study proposes a residual-based statistic for validating the Q-matrix. Its performance is evaluated in a simulation study and compared against the method of Liu, Xu, and Ying (2012). Simulation results indicate that the proposed method leads to a higher recovery rate of the Q-matrix and requires less computational time.

Nonparametric Attribute Profile Estimation and Q-Matrix Reconstruction Using Modified Auto-Encoder

Kang Xue, University of Georgia; Laine Bradshaw, University of Georgia - Athens

The goal of this research was to estimate students' attribute profiles without specific probabilistic models and to reconstruct an inaccurate Q-matrix. In this paper, a modified autoencoder network was designed to achieve the research task. Simulated experiments were conducted to test the performance of our method under different assessment conditions.

Investigation of the Model Invariance for Diagnostic Classification Model With Polytomous Attributes

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

Most diagnostic classification models (DCMs) provide dichotomous feedback about students' mastery and non-mastery levels. The DCM for polytomous attributes (PDCM) can classify students into more than two mastery levels. We examined the group invariance and item invariance property for the PDCM by conducting a simulation study and an empirical study.

Sunday, April 7, 2019
12:10 – 1:40pm, Alberta, Coordinated Session

Communicating Achievement Results That Incorporate Response Time Data: Challenges and Advances

Chair: Jesper Tijmstra, Tilburg University
Discussant: Jesper Tijmstra, Tilburg University

With the advance of computerized testing in educational measurement, it has become commonplace to record response time (RT) in addition to response accuracy. Psychometric models have been developed that take both the speed and correctness of responses into account. These models generally have two benefits: Ability is estimated with a greater precision, and a more complete picture of the performance of the respondent is obtained. Because these models are also simple in structure, they can efficiently summarize the performance of persons or groups on the test and hence can be helpful for communicating to stakeholders. Two important issues have to be considered when assessing whether RT models should be used for summarizing test performance: (1) Commonly used models are simple in structure and are likely to be misspecified; (2) These models are likely not utilizing all relevant information available in the response time data. Both these issues are important, as they may prevent standard response time models from being optimal tools for summarizing and communicating test results in practice. The proposed session consists of four presentations that all focus on these two connected issues and that present innovative ways of utilizing information in the RT data.

Increasing Precision Is Not Everything: The Impact of Disengagement on Inferences Under the Hierarchical Model for Response Time and Accuracy
Maria Bolsinova, ACT, Inc.; Jesper Tijmstra, Tilburg University

Accounting for Individual Differences in Speed in the Discretized Signed Residual Time Model
Jesper Tijmstra, Tilburg University; Maria Bolsinova, ACT, Inc.

Response Time Processes in Computerized Adaptive Testing
Yang Shi, University of California - Berkeley; Kyung (Chris) T. Han, The Graduate Management Admission Council

Roles and Uses of Response Times in Psychometrics: An Example of Studying Automated Knowledge Retrieval Cognitive Process
Paul De Boeck, The Ohio State University; Minjeong Jeon, University of California at Los Angeles

Sunday, April 7, 2019
12:10 – 1:40pm, Algonquin, Coordinated Session

Comparing Automated Scores With Human Scores of Essays in Writing Assessments

Chair: Wei Wang, Educational Testing Service
Discussant: Mark Shermis, University of Houston-Clear Lake

In recent years, automated essay scoring (as opposed to human scoring) has been increasingly used in various kinds of standardized tests. This coordinated session discusses several topics regarding automated essay scoring. The first paper provides an introduction of e-rater®, ETS' automated essay scoring engine, and its current uses in three high-stakes standardized tests. The second paper provides in-depth discussion on when automated scores on essay writing, which are predictions of human scores, can serve as a substitute for human scores. The third paper compares agreement and prediction statistics as measures of the performance of automated essay scoring. The fourth paper compares the prediction of writing true scores using scaled vs. non-scaled automated scores.

Current Uses of e-rater® Automated Essay Scoring in High-Stakes Assessments
Mo Zhang, Educational Testing Service

Automated Scores of Writing Performance: Predictions of Human Scores or Substitutes for Human Scores?
Neil J. Dorans, ETS

Comparing Agreement and Prediction Statistics as Measures of the Performance of Automated Essay Scoring
Wei Wang, Educational Testing Service; Neil J. Dorans, ETS

Investigating the Effects of Scaling in Predicting Human True Scores
Mo Zhang, Educational Testing Service; Wei Wang, Educational Testing Service; Neil J. Dorans, ETS; Chen Li, ETS; Lili Yao, Educational Testing Service

Sunday, April 7, 2019
12:10 – 1:40pm, Ballroom, Coordinated Session

Communicating Assessment Results: How to Inform Decision Making in Education

Chair: Okan Bulut, University of Alberta
Discussant: Michael Jodoin, National Board of Medical Examiners

In today's ever-changing technology landscape, education also continues to evolve and adopt new technologies that are expected to help major stakeholders make better decisions. Despite the availability of data from many promising applications of assessments, more data do not necessarily guarantee better decisions. In fact, massive amounts of information from assessments might result in information overload for individuals and confusion about assessment practices. Therefore, there is still a need for innovative and yet simple applications that not only communicate assessment results to major stakeholders in education, but also help them utilize the information for making better decisions. This session includes four presentations about innovative assessment practices that can help students, instructors, and testing programs to make informed decisions. The first presentation demonstrates how to utilize process data from a learning management system to identify students who might be at risk of failing a course. The second presentation focuses on the use of score reporting for understanding students' achievement goals and promoting their learning. The third presentation introduces a data-driven method for extracting students' misconceptions from written responses to inform instructors and test developers. The last presentation demonstrates a score reporting system that communicates assessment results to via interactive visualizations.

Using Learning Management Data to Predict Student Course Performance
Ying Cui, University of Alberta

Immediate Score Reporting: Pausing to Consider the Motivational and Emotional Consequences for Students
Lia Marie Daniels, University of Alberta; Okan Bulut, University of Alberta; Mark J. Gierl, University of Alberta

Using Students' Written Responses to Inform Content Specialists About Students' Common Misconceptions
Jinnie Shin, University of Alberta; Qi Guo, University of Alberta; Mark J. Gierl, University of Alberta

ExamVis: A Score Reporting System for Visual Communication of Test Results
Okan Bulut, University of Alberta; Maria Cutumisu, University of Alberta

Sunday, April 7, 2019
12:10 – 1:40pm, British Columbia, Coordinated Session

Computational Psychometrics for Learning and Assessment in Virtual Environments

Chair: Alina A. Von Davier, ACT, Inc.

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 big data collected from virtual environments should be intentional: we should 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. CP is also interested in the knowledge discovery from the (big) data (KDD). In this symposium, several examples of applications of computational models for learning and assessment are presented. Psychometrics theories and data-driven algorithms are fused to make accurate and valid inferences in complex, virtual learning and assessment environments.

Computational Psychometrics — Hype or Hope?
Alina A. Von Davier, ACT, Inc.

Beyond Assessment: A Computational Psychometrics Approach to Foster New Way of Measurements in Education
Pietro Cipresso, Applied Technology for Neuro-Psychology Lab, Catholic University of Milan

The Wiring of Intelligence (and Other Latent Variables)
Gunter Maris, ACTNext

Heterogeneous Effects of Adaptive Tutoring on Grades and Major Choice: A Quasi-Experiment
Thomas Fikes, Arizona State University; Rene kizilcec, Cornell University

Targeting the Data and Modeling Challenges From Next-Generation Assessments: A Computational Perspective on Psychometrics
Jiangang Hao, ETS

Sunday, April 7, 2019
12:10 – 1:40pm, Quebec, Paper Session

Important Considerations in Computerized Adaptive Testing and Item Pool Utilization

Discussant: Kristin M. Morrison, ACT, Inc.

Item Pool Utilization Based on Item Selection Methods and Test Termination Rules

Sema Sulak, Bartin University

The purpose of this study is to determine item selection methods for item pool utilization. The results of this study suggest that the performance of item selection methods with regard to item pool utilization highly depends on the test termination rule as well as the ability estimation method.

Design and Comparison of Four Stopping Rules in Mastery Computerized Adaptive Testing and On-the-Fly Multistage Testing (OMST)

Chen Tian; Hua-Hua Chang, Purdue University

The stopping rule plays a critical role in mastery testing. In this study, we proposed a new truncation rule designed particularly for adaptive testing, explored and compared the efficiency of different stopping rules, and examined their application to on-the-fly multistage testing.

Routing Strategies and Optimizing Design for Multistage Testing in International Large-Scale Assessments

Dubravka Svetina, Indiana University - Bloomington; Yuan-Ling Liaw, University of Oslo; Leslie Rutkowski, Indiana University; David Rutkowski, Indiana University

Acknowledging the advantages of multistage testing (MST), in 2013, PIAAC was the first international large-scale assessment (ILSA) to incorporate such design. Our simulation study investigates conditions under which item exposure is maximized and thetas are well recovered, both of which are relevant yet largely unexplored in ILSA context.

Controlling Test Speededness in Computerized Adaptive Testing

Zhuoran Wang, University of Minnesota - Twin Cities; Edison M. Choe, The Graduate Management Admission Council

New item selection criteria for CAT were developed to control for test speededness by reducing both the number of unfinished examinees and variability of test completion time. A simulation study based on real data demonstrated the efficacy of the methods while balancing ability estimation accuracy and item pool usage.

Controlling Minimum Item Exposure Rates on the Fly in Computerized Adaptive Testing

Jyun-Hong Chen, Soochow University; Hsiu-Yi Chao, National Taiwan University

A procedure for controlling minimum item exposure rate in CAT was proposed. Items with exposure rates lower than the pre-specified value are made to increase their usage in early stages. Simulation results indicated that the procedure can efficiently improve the item usage while maintaining the precision level of trait estimates.

Sunday, April 7, 2019
12:10 – 1:40pm, Salon A, Paper Session

Investigating Student Growth and Learning

Discussant: Jonathan P. Weeks, Educational Testing Service

No Pain, No Gain: Lesson Learned From a Test Prep Experiment

Edgar I. Sanchez, ACT, Inc.; Ty M. Cruce, ACT, Inc.

Many high school students turn to test-preparation programs to help them improve their chances of college admissions. We used an experimental design to evaluate the effectiveness of an online test-prep course to improve test scores. Most students did not use the course as intended, resulting in no score gains.

Investigating the Relationship Between Test Preparation Activities and Students' ACT Composite Scores.

Raeal Moore, ACT, Inc.; Edgar I. Sanchez, ACT, Inc.

A study was conducted to determine the relationship between types of test preparation activities and students' ACT Composite scores. Emphasis was placed on whether this relationship depended on students' family income and times taking the test. Findings are used to explore fairness issues in admissions testing.

Analyzing Learning Processes and Distinct Learning Patterns in Higher Education Economics

Jasmin Schlax, Johannes Gutenberg-University; Olga Zlatkin-Troitschanskaia, Johannes Gutenberg University of Mainz; Susanne Schmidt, Johannes Gutenberg-Universität Mainz; Carla Kühling-Thees, Johannes Gutenberg-University Mainz; Judith Jitomirski, Humboldt University - Berlin; Roland Happ, Johannes Gutenberg University of Mainz

The aim of higher education is to increase students' knowledge. In a longitudinal panel sample of 748 students of economics, we found different patterns of knowledge development and learning (positive, negative, retained and zero), which provide us with much more differentiated insight into growth than the overall knowledge test scores.

Applying Gradient Boosted Regression Trees to Produce Growth Percentiles

Steven Tang, eMetric; Zhen Li, eMetric

This study analyzes the potential of using a gradient boosted regression tree (GBRT) model to compute growth percentile rankings in summative accountability contexts. Results indicate that with default hyperparameters, GBRT percentile rank residuals replicate standard SGP and that GBRT may be further tuned to potentially predict more accurately.

Item Response Theory Modeling of Decomposed Student Learning Patterns in Higher Education Economics

Susanne Schmidt, Johannes Gutenberg-Universität Mainz; Olga Zlatkin-Troitschanskaia, Johannes Gutenberg University of Mainz; William B. Walstad, University of Nebraska

This study describes a new sophisticated modeling approach for differentiated analyses of aggregated test scores from a multiple-choice test in economics 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.

Sunday, April 7, 2019
12:10 – 1:40pm, Salon B, Paper Session

Measurement and Policies Surrounding Accountability Testing

Discussant: Fran Stancavage, American Institutes for Research

Discontinuities and Unbiased Rescoring Policies in High School Exit Exams

Sophie Litschwartz, Harvard University

In 2011, the Wall Street Journal reported that teacher score manipulation caused discontinuities in the NYC Regent test score distribution around the pass cutoff. In this paper, I show how test re-scoring and even a small amount of test error could have caused a discontinuity without any teacher manipulation.

Measuring Instruction With E-Portfolios: Reliability With Instructional Units of Different Lengths

Jose Felipe Martinez, University of California - Los Angeles; Jayashri Srinivasan, University of California - Los Angeles; Matthew J. Kloser, University of Notre Dame; Brian Stecher, The RAND Corporation; Amanda Edelman, Pardee RAND Graduate School

Teacher portfolios hold promise for monitoring and improving instruction at scale. We examine the properties of measures of instruction derived using a new type of e-portfolio tool for mobile devices. We find comparable levels of reliability with measures obtained using portfolios that cover one and two weeks of instruction.

Evaluating the Persuasiveness of Policy on 11+ Testing in Trinidad and Tobago

Jerome De Lisle, University of the West Indies

Multiple tools were used to explicate and evaluate historic and current policy arguments for high stakes 11+ testing in Trinidad and Tobago. Stakeholder evaluation suggested that some respondents were persuaded by specific assessment designs and issues. An evolving, dynamic system possibly contributes to continued public legitimacy despite unintended negative consequences.

Models of Using College Entrance Examinations for Accountability

Michelle Croft, ACT, Inc.; Gretchen Guffy, ACT, Inc.; Dan Vitale, ACT, Inc.

States are including college entrance examination scores such as the ACT/SAT for federal accountability. This paper provides case studies of how states are incorporating the ACT/SAT and identifies areas where the types of validity evidence needed to support the use of the assessment may differ depending on the use.

Sunday, April 7, 2019
12:10 – 1:40pm, Territories, Paper Session

New Directions in Item Response Theory

Discussant: Scott Monroe, University of Massachusetts - Amherst

Pairwise Comparison Using a Bayesian Selection Algorithm: Efficient Holistic Measurement

Elise Crompvoets, Tilburg University; Anton Beguin, Cito; Klaas Sijtsma, Tilburg University

Pairwise comparison is becoming an increasingly popular assessment method. Unfortunately, many comparisons are required for reliable measurement. Adaptive pairwise comparison seems promising to reduce the required number of comparisons, but current algorithms are suboptimal. We proposed a Bayesian algorithm as a solution and tested its performance in a simulation study.

A Multilevel Mixture Item Response Theory Framework for Modeling Guessing Behavior in Proficiency Tests

Gabriel Nagy, Leibniz Institute for Science and Mathematics Education; Alexander Robitzsch, IPN

A multilevel mixture IRT framework for modelling guessing on the item level is presented. Guessing can be specifi as a function of item and person characteristics, which means that different IRT models fi into the framework. The utility of the framework is demonstrated on the basis of a reading test.

Using the Discontinuation Rule to Reduce the Effect of Random Guessing

Tianshu Pan, Pearson; Youngmi Cho, American Institutes for Research

In this article, a Monte Carlo study was implemented to explore the feasibility of the discontinuation rules to reduce the effect of random guessing on item-parameter estimation in the Rasch model. The results showed the discontinuation rules can reduce this effect for item- parameter estimation.

Some Programming Techniques in Stan for Advanced Item Response Theory Models

Shaoyang Guo, East China Normal University; Chanjin Zheng, Jiangxi Normal University

This study recommends two programming techniques in Stan, vectorization and the combination of built-in functions, which could obviously accelerate computational effi y. To illustrate the advantages of these techniques, four real-world datasets with different IRT models were selected as examples for the dichotomous and polytomous IRT models, compared with OpenBUGS.

Sunday, April 7, 2019
3:20 – 4:50pm, Alberta, Invited Speaker Session

2019 NCME Career Award Session
Chair: Neil J. Dorans, ETS
Discussant: Sandip Sinharay, Educational Testing Service

Statistical Theory and Assessment Practice

Shelby Haberman

Sunday, April 7, 2019
3:20–4:50pm, Algonquin, Coordinated Session

Do Medical Licensing/Certification Exams Really Make a Difference?

Chair: Liane N. Patsula, Medical Council of Canada
Discussant: Michael T. Kane, ETS

The primary use of medical licensure/certifi tion examinations is to assure the public that a candidate has adequate knowledge and skills necessary for safe and effective patient care. However, there is growing interest in assessing whether licensing/certification examination results can also be used for additional purposes such as predicting physician practice performance. Focusing on the extrapolation and implications inferences of Kane's (1992) validity argument, the studies in this session are aimed at gathering validity evidence to support potential defensible secondary uses of examination results and evaluating the impact of a new national licensing examination. • The first paper examines relationships between United States Medical Licensing Examination scores and formal disciplinary sanctions physicians receive in practice. • The second and third papers assess whether there is a predictive relationship between physician performance on the Medical Council of Canada Qualifying Examinations and post-licensure peer assessment for physicians in Ontario and patient complaints, and opioid and benzodiazepine prescription patterns for physicians in Alberta. • The fourth paper studies the association between maintaining certification in general surgery and loss of license actions over a 30-year period. • The last paper evaluates the impact of a new national licensing examination on faculty and students across 18 medical schools in Indonesia.

Evaluating Validity Evidence for a Medical Licensure Assessment System With Important Secondary Uses Through a Multilevel Examination of External Professional Practice Outcomes
Monica M. Cuddy, National Board of Medical Examiners

Validity Extrapolation Beyond Licensure: Using Physician Performance on Licensing Examinations to Predict Their Performance in Practice
Fang Tian, Medical Council of Canada; André F. De Champlain, Medical Council of Canada; Wendy Yen, College of Physicians and Surgeons of Ontario; Niels Thakker, College of Physicians and Surgeons of Ontario; Dan Faulkner, College of Physicians and Surgeons of Ontario; Sirius Qin, Medical Council of Canada

Do National Licensing Examination Scores Predict Patient Complaints and Physician Prescribing Patterns? Gathering Secondary Validity Evidence for a Large-Scale Medical Licensing Examination Program
André F. De Champlain, Medical Council of Canada; Nigel Ashworth, College of Physicians and Surgeons of Alberta; Sirius Qin, Medical Council of Canada; Delaney Wiebe, College of Physicians and Surgeons of Alberta; Nicole Kain, College of Physicians and Surgeons of Alberta

Association Between Maintaining Certification in General Surgery and Loss of License Actions
Andrew Jones, American Board of Surgery; Jason P. Kopp, American Board of Surgery; Mark Malangoni, American Board of Surgery

Beyond Test Scores: Researching the Outcomes of a Recently Introduced National Licensing Exam on Medical Schools in Indonesia
Trudie Elizabeth Roberts, University of Leeds; Rachmadya Nur Hidayah, University of Leeds; Richard Fuller, University of Leeds

Sunday, April 7, 2019
3:20 – 4:50pm, Ballroom, Paper Session

Issues and Advances in Automated Scoring of Constructed-Response Items

Discussant: Ada Woo, ACT Inc

Communicating to the Public About Machine Scoring: What Works, What Doesn't

Mark D. Shermis, University of Houston--Clear Lake; Sue Lottridge, American Institutes of Research

This paper documents six case studies about how to, and how not to, communicate a k-12 testing entity's transition to machine scoring. Data are drawn from four U.S. state-, one Canadian Province-, and one country's testing programs. Based on the analysis of the six cases, several tentative recommendations are made.

Human Rating Errors and the Training of Automated Raters

Richard J. Patz, University of California Berkeley; Sue Lottridge, American Institutes of Research; Michelle Boyer, Data Recognition Corporation

Human raters are subject to defects of inconsistency and bias. How such defects among individual raters impact overall measures of rating quality and the accuracy of automated ratings is studied. Opportunities for the modeling of rater characteristics to improve the training of automated raters are examined.

The Impact of Automated Scoring of Essay Writing on Reporting Scores

Youngmi Cho, American Institutes for Research; Sue Lottridge, American Institutes of Research; Ahmet Turhan, American Institutes for Research

This study investigates the impact of the performance of automated scoring evaluated by agreement statistics and standardized mean difference between automated and human scores at reporting-level scores. Results from real and simulated data presenting various rater's behaviors provide guidelines to ensure the validity of using automated scores in high-stakes assessment.

Stop Word Selection for Latent Dirichlet Analysis of Constructed-Response Items

Minho Kwak, University of Georgia - Athens; Seohyun Kim, University of Georgia; Jiawe Xiong, University of Georgia; Choi Hyejung, University of Georgia; Allan S. Cohen, University of Georgia

This study investigated the impact of TF-IDF stopword selection method on the LDA analysis of constructed response items. LDA extracts latent topics from textual data. The results indicated that the impact of the removal of stopwords was negligible when the number of the stopword removed was less than 50.

Toward Improving the Machine Scorability of Constructed-Response Items

Allan S. Cohen, University of Georgia; Kevin Raczynski, University of Georgia; Holly Garner, University of Georgia

Machine scoring is increasingly being used to grade responses to constructed response items. Characteristics of items that are, or are not, machine scorable have not been widely studied. We gather empirical evidence for whether items are machine scorable and then describe characteristics of items that are, or are not.

Sunday, April 7, 2019
3:20 – 4:50pm, British Columbia, Paper Session

Score Comparability: Matters of Mode

Discussant: Richard M. Luecht, University of North Carolina - Greensboro

Investigating the Impact of Mode on Students' Response Processes in Educational Assessment

Yile Zhou, The University of Iowa; Justin Paulsen, Indiana University; Kristin M. Morrison, ACT, Inc.

A think-aloud study was conducted with a repeated-measure design to evaluate the impact of mode (paper vs. computer) on students' response processes. Students were presented with multiple-choice items in both administration modes. A coding scheme is developed and applied to analyze students' behaviors.

Rater Effect Study in Dual Mode Testing: Hierarchical Rater Model Approach

Kyoungwon Lee Bishop, WIDA at University of Wisconsin -Madison

In this study, we explore the rater effects of test administrative modes (Paper vs. Online) of the Speaking domain in ACCESS for ELLs. We address the mode effect by modeling ratings by Heirarchical Rater Model and also with observations on scoring, interviews with raters, and discourse analysis on spoken responses.

Does Digital Familiarity Affect Performance in NAEP's 2012 Digitally Based Writing Assessment?

Young Yee Kim, American Institutes for Research; Xiaying Zheng, American Institutes for Research; Youmi Suk, The University of Wisconsin - Madison

There has been a concern that fourth-graders' digital familiarity may not allow them to fully demonstrate their writing ability in digitally based assessment (DBA) in NAEP. The purpose of this study is to investigate the relationship between writing performance and digital-familiarity related factors using NAEP 2017 grade-4 DBA writing assessment.

Using Response Process Data to Examine Comparability Across Writing Prompt Interfaces

Justin Paulsen, Indiana University; Kristin M. Morrison, ACT, Inc.

An eye-tracking study was conducted to examine comparability of response processes across two different computer interfaces. Students received two equivalent but different writing prompts in the two interfaces and responded to the items. Fixation data and qualitative descriptions of the gaze patterns were used to analyze differences across interfaces.

Item- and Test-Level Statistical Adjustment Across Test Administration Modes

Nina Deng, Questar Assessment, Inc; Quintin Ulysses Love, Questar Assessment Inc.; Katherine Nolan, Questar Assessment Inc.

It becomes common that statistical adjustments are made when multiple test administration modes are offered. This paper evaluates the impacts of mode adjustments on students' test score and performance classifi tion using real and simulated data. Various item- and test-level adjustments are compared and their relationships are investigated under different conditions.

Sunday, April 7, 2019
3:20 – 4:50pm, Manitoba, Paper Session

Advances in Differential Item Functioning Detection and Research

Discussant: Michelle Boyer, Data Recognition Corporation

Evaluating Bayesian Diff ential Item Functioning Detection Methods Using an Ideal Point Item Response Theory Model

Seang-Hwane Joo, KU Leuven; Phil Seok Lee, George Mason University; Stephen E. Stark, University of South Florida

We examined the performance of two Bayesian DIF detection methods, Bayes Factor (BF) and Deviance Information Criterion (DIC), using the ideal point IRT model GGUM. Power and Type I error were investigated via a Monte Carlo study. We will present the results and provide recommendations for test developers and practitioners.

Differential Item Functioning Detection and Parameter Recovery Using the C-RUM Model

Kevin Krost, Virginia Polytechnic Institute and State University; Gary E. Skaggs, Virginia Polytechnic Institute and State University

This simulation study evaluated parameter recovery and compared power and Type I errors for DIF detection between the Wald test and likelihood ratio test, using the compensatory reparameterized unified model. Sample size, DIF type, magnitude, and Q-matrix complexity were factors which affected each outcome. Implications discussed.

Detecting Differential Item Functioning Using the Adaptive LASSO Penalty

Jing Jiang, Boston College; Zhushan Mandy Li, Boston College

This paper examinees the applicability of using regularization methods in estimation for simultaneous DIF detection of all items on a test, and introduces the use of the adaptive LASSO penalty, which is compared with the standard LASSO penalty through a comprehensive simulation study.

Examining Differential Item Functioning Using Projective Item Response Theory Modeling

Terry A. Ackerman, University of Iowa; Ye Ma, The University of Iowa; Jinmin Chung, The University of Iowa

The purpose of this study is to investigate the potential of the Projective IRT Model to eliminate or ameliorate differential item functioning in dichotomously scored tests. This study includes both real and simulated data analyses in which DIF is examined using Raju's area method, SIBTEST, and Mantel-Haenszel procedures.

A Differential Item Response Model for the Effect of a Continuous Person Covariate of Vocabulary Knowledge

Saemi Park, The Ohio State University - Columbus; Paul De Boeck, The Ohio State University

We propose a new framework for DIF with a continuous person covariate. We applied it to a vocabulary test using a reading comprehension score a covariate to explore item slopes as a function of reading comprehension and found that strong readers perform better for more polysemous and frequent the words.

Sunday, April 7, 2019
3:20 – 4:50pm, Quebec, Paper Session

Technical Considerations in Item Response Theory

Discussant: William Skorupski, Amira Learning

A Bayesian Two-Tier Item Response Theory Model Suitable for Small Samples

Ken A. Fujimoto, Loyola University Chicago

This presentation includes details of a Bayesian version of the two-tier item response theory model. This model uses prior distributions that makes it applicable to data from smaller and larger samples (e.g., 100 and 1,000 individuals, respectively). Simulated and empirical data are analyzed to demonstrate this model's performance.

Using Expectation Maximization (EM) for Finite Mixtures and Supplemented EM to Analyze Multidimensional Item Response Theory Data

Ping Chen, Beijing Normal University; Chun Wang, University of Washington

This study revisits the parameter estimation issues in MIRT more deeply and investigates some computation details that haven't been addressed in implementing EM for finite mixtures, e.g., rescale after each EM cycle or after final EM cycle? How to apply supplemented EM to estimate standard errors of all unknown parameters?

Standard Error Adjustment for Projected Item Response Theory Scores in a Fixed-Item-Parameter Linking Design

Shuangshuang Xu, University of Maryland - College Park; Shuangshuang Xu, University of Maryland - College Park; Yang Liu, University of Maryland - College Park

To account for the carry-over sampling variability to projected scores in IRT, we adjust standard errors using multiple imputation. Imputed parameter sets come from a multivariate normal approximation to their sampling distribution, in which the mean vector and the asymptotic covariance matrix are obtained from two-stage pseudo-maximum likelihood estimation.

A Two-Decision Unfolding Tree Model for Likert Scale Items

Kuan-Yu Jin, The University of Hong Kong; Hui-Fang Chen, City University of Hong Kong; Yi-jhen Wu, Florida State University

IRT-tree models have recently become popular in analyzing rating-scaled data. We proposed a two-stage IRT-tree model by incorporating an unfolding model to a multi-process tree model in order to accurately refl t the cognitive process of endorsing a response category. The new model was applied to European Social Survey 2016.

Sunday, April 7, 2019
3:20 – 4:50pm, Salon A, Coordinated Session

The Estimation and Scaling of Rater Effects Parameters for Large-Scale Rater Monitoring

Chair: Jodi M. Casabianca, Educational Testing Service
Discussant: Stefanie Wind, The University of Alabama
Discussant: Won-Chan Lee, University of Iowa

Rater response models may be used to quantify the extent to which raters exhibit errors in their scoring. Recently, researchers at ETS have focused on using this modeling approach to develop a comprehensive database of rater effects indices for large-scale testing programs which involve several hundred raters, scoring hundreds of items, sometimes on a weekly basis. The papers in this session focus on this modeling approach and the challenges involved in creating raters effects indices on the same scale. Casabianca introduces the methods under consideration, including the estimation and scaling techniques, and reports on best practices derived from a simulation study and empirical data analyses. Shin focuses on an alternative technique for scaling rater effects parameters across multiple items—using a multiple-group IRT model. Donoghue addresses the extent to which the indices captured describing rater behavior are mostly fixed or if there are interactions with the item yielding a larger amount of variability in behavior. Choi examines test taker population differences and determines if there is differential rater functioning related to native language of the response, impacting summaries of rater effects indices from different test forms. All papers focus on the same datasets from an international assessment of English language.

Techniques and Assumptions for Rater Response Model Estimation and Rater Scale Linking
Jodi M. Casabianca, Educational Testing Service

Estimating Rater Effects Across Multiple Items Using Multiple-Group Item Response Theory Models
Hyo Jeong Shin, Educational Testing Service

Examining Stability of Rater Effects Across Items
John R. Donoghue, Educational Testing Service

Investigating Human Rater Bias With a Sparse Rater Assignment Matrix
Ikkyu Choi, Educational Testing Service

Sunday, April 7, 2019
3:20 – 4:50pm, Salon B, Paper Session

Lost and Found: Techniques for Handling Missing Data

Discussant: Thanos Patelis, Human Resources Research Organization

Effects of Changing Nonresponse Mechanisms on Trends and Comparisons in Large-Scale Assessments

Karoline A. Sachse, Humboldt University - Berlin; Nicole Haag, Humboldt University - Berlin; Steffi Pohl, Free University - Berlin

We reanalyzed PISA data of three cycles and found considerable variation in nonresponse rates and mechanisms. A simulation study based on the results showed severely biased cross-sectional trend estimates if changing mechanisms are not taken into account. We evaluated different missing data approaches. Practical implications of the results are discussed.

A Modified Algorithm for the Use of Plausible Values in Assessment Surveys

Andrew J. Kolstad, P20 Strategies LLC

One normally calculates standard errors from U (standard error of a set of PVs) and B (variance among PV set means). My alternative estimates B from average variance of examinee PVs, effectively increasing B's sample. I evaluate the modification with 300 PVs from the 2003 National Assessment of Adult Literacy.

Nonresponse Issue in Noncognitive Measures: Validity Approach Using Explanatory Item Response Modeling

Jiaying Xiao, University of Alberta; Okan Bulut, University of Alberta; Michael C. Rodriguez, University of Minnesota

Using explanatory item response modeling, we examine validation concerns due to nonresponse issue in a noncognitive measure of bullying. We find that significant interactions between gender, grade levels, and type of bullying behavior measured in the items lead to different nonresponse patterns and thus potentially influence score interpretation.

Correcting for Student Dropout in Longitudinal Item Response Theory Assessment Using Commercial Software

Charles J. Iaconangelo, Pharmerit International; Daniel Serrano, Pharmerit International

There is substantial interest in tracking student achievement over time. However, student drop-out threatens the validity of inferences made using this data. This research proposes a method for correcting for missing not at random (MNAR) in longitudinal IRT analysis. This approach is straightforward to implement in existing commercial software.

Sunday, April 7, 2019
3:20 – 4:50pm, Territories, Invited Speaker Session

The Influence of Stakeholder Needs and Values on Assessment Design and Reporting

Discussant: Kristen L. Huff, Curriculum Associates, Inc.
Chair: Paul D. Nichols, ACT, Inc.

Different stakeholders in a testing program will bring different conventions, practices, values, and needs to interactions with that testing program. For example, research suggests the values and needs of psychometricians tend to emphasize: • The influence of behaviorism; • The use of quantitative methods and arguments for the validity of inferences made from assessment results and the generalizability of findings; • Achievement as located in the learner and conceptualized as a latent variable; In contrast, the values and needs of teachers tend to be shaped by their location in the education hierarchy and the kinds of responsibilities assigned to their positions. Teachers tend to emphasize: • The importance of local and situational factors; • The value of evidence that captures student thinking and that is rooted in authentic classroom contexts; and, • The use of clinical judgment; The implication for a testing program is that the artefacts produced by a testing program and intended to communicate with different stakeholders should be designed and developed to account for the different values and needs of each stakeholder group. In this session, we explore the implication of stakeholder values and needs for three artefacts produced by a testing program: validity arguments, score reports, and technical reports.

A Framework for Understanding Assessment Stakeholders' Values and Needs
Paul D. Nichols, ACT, Inc.

A Principled Approach to Score Reporting in Support of Users' Needs and Values
Daniel Lewis, ACT

Who Cares About Technical Reports and Why?
Jeff Michael Allen, ACT, Inc.

Test Design, Score Reporting, and Validity Arguments for Different Audiences: Growth Reporting as Illustration
Steve Ferrara, Measured Progress

Sunday, April 7, 2019
5:05 – 6:35pm, Alberta, Paper Session

New Learning in Item Analysis Research

Discussant: Hollis Lai, University of Alberta

Three Methods of Item Analysis

Seock-Ho Kim, University of Georgia; Allan S. Cohen, University of Georgia; Hyo Jin Eom, University of Georgia

This paper contrasts three methods of item analysis based on classical test theory, generalized linear modeling, and item response theory. Illustrations of the methods are presented with simulated and real data. Specifically, the methods respectively use a cross classification table under classical test theory, a baseline-category logit model under generalized linear modeling, and a multiple choice model under item response theory. Advantages and disadvantages of each method are discussed.

Pretest Item Calibration in Multistage Adaptive Testing

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

This study aims to evaluate methods for calibrating and linking pretest items in a 1-3 MST design when the pretest items are administered together with operational items. Performance of fixed parameter calibration and separate calibration with linking methods are compared in terms of item parameter recovery using simulated data.

Positive Intercultural Adaptation: Item Weighting and Differential Item Functioning

Travis Henry, University of Georgia - Athens; Pedro R. Portes, University of Georgia; Ruben Atilano, University of Georgia - Athens; Diego Boada Beltran, University of Georgia - Athens

Positive Inter-cultural Adaptation measures successful acculturation (Authors, 2016). Undergraduate data (N = 3,491) from one Hispanic serving university and one historically white university were analyzed using a three-facet Rasch model. Results indicate no DIF for university. On average, students from both universities form inter-cultural identities with about the same ease.

Anchors Aweigh: How the Choice of Anchor Items Affects Rasch Vertical Scaling

Tom Waterbury, James Madison University; Christine Demars, James Madison University

Vertical scales were constructed using Rasch modeling with simulated 3PL item responses. Scaling was conducted with either relatively easy or difficult anchor items. While the presence of correct guessing biased growth estimates regardless of anchor item choice, growth was much more severely underestimated when using the difficult anchor items.

Randomly Clicking on Experimental Items and Item Parameter Estimation

Xiaoliang Zhou, Teachers College, Columbia University

The present study aims to explore how randomly clicking on experimental items may affect item parameter estimation in IRT models. Using a simulation study, it was found that increasing the proportion of randomly clicking examinees increased the bias and RMSE for both discrimination and difficulty parameter estimations.

Sunday, April 7, 2019
5:05 – 6:35pm, Algonquin, Coordinated Session

Useful and Usable Learning Analytics
Chair: David Michael Niemi, Kaplan
Discussant: Roy D. Pea, Stanford University

The digital revolution in education has dramatically increased the streams of data available for interpretation, analysis, and visualization. Not only are data plentiful, they are increasingly inexpensive to collect and analyze. Whether educators and learners can use the expanding ocean of data to improve learning, however, depends to a great extent on whether the relatively new field of Learning Analytics can be organized to focus on: 1) providing useful information to those who can act on it (including students) and 2) making sure that users know what to do with the information. Consistent with these aims and the Annual Meeting theme, this panel discussion session will open with a group of expert panelists giving brief talks on examples of different successful uses of analytics, including providing data to students, using machine learning to evaluate multimodal student artifacts, using data mining to recommend interventions, and diagnosing and intervening to enhance motivation and persistence. These introductory comments are designed to jump-start an open-ended panel-and-audience discussion of issues raised by the panelists—including whether and how Learning Analytics can be used to improve equity for learning and career success for all students—as well as questions posed by the audience.

Putting the Learner at the Center: Sharing Analytics With Learning Participants
Marie Bienkowski, SRI International

Multimodal Learning Analytics and Assessment of Open-Ended Artifacts
Paulo Blikstein, Stanford University

Demonstrating the Value of Educational Data Mining
Ryan Shaun Baker, University of Pennsylvania; Kenneth R. Koedinger, Carnegie Mellon University

Analytics to Improve Motivation and Persistence
Bror Valdemar Haug Saxberg, Chan Zuckerberg Initative; Richard Clark, University of Southern California

Sunday, April 7, 2019
5:05 – 6:35pm, British Columbia, Coordinated Session

What About Psychometrics in Formative Assessments?

Chair: Saskia Wools, Cito
Discussant: Bernard P. Veldkamp, Universiteit Twente

In formative assessment, evidence about students' performance is used to make decisions about instructional actions. Although these decisions are less life-changing then some summative tests, these assessments still have to adhere to basic psychometric criteria. Unfortunately, the low stakes character of these formative assessments and the requirement to make them efficient within an everyday classroom setting, provides us with psychometric challenges. To understand formative assessment from a psychometric perspective, it is necessary to re-evaluate psychometric concepts. Do concepts such as validity, reliability, standard setting and test designs remain the same when used for formative assessments? And if not, how would we operationalize and use them in a way that ensure the decision quality within a formative assessment context? In this coordinated session, three papers are presented to answer these questions. All papers are tied to a formative assessment platform (GM) for math in K12 education. The first paper addresses the platform and draws conclusions about the design of formative assessments. The second paper presents a framework for validation of formative assessments. The final paper aims to establish the optimal test length and cut-scores for formative assessments within the GM platform.

General Principles for Formative Assessment: Results of a Design-Based Research Study
Saskia Wools, Cito

A General Framework for the Validation of Embedded Formative Assessment
Dorien Hopster-den Otter, University of Twente

On Effective Test Lengths and Cutoff Scores in Learning Objective–Based Mastery Tests
Hendrik Straat, Cito

Sunday, April 7, 2019
5:05 – 6:35pm, Imperial Room, Electronic Board Session

Electronic Board Session 3

Can Simultaneous Linking Reduce Item Parameter Drift on Vertical Scales?
Lixiong Gu, ETS; Jiyun Zu, ETS; Longjuan Liang, ETS

Typical vertical scales are maintained by on-level horizontal equating only after the scale is initially established. This simulation study explores whether simultaneous linking method, which estimates equating constants simultaneously for test forms of different levels (vertically) and administrations (horizontally), can reduce item parameter drift on vertical scales.

Comparing Joint Versus Separate Calibration of Multiple Tests With a Generalized Diagnostic Classification Model
Yanhong Bian, Rutgers, the State University of New Jersey; Benjamin R. Shear, University of Colorado - Boulder; Louis A. Roussos, Measured Progress

This study explores how classification consistency for three diagnostic geometry tests is affected by calibrating the tests separately or jointly in a single model. Because each test diagnoses both skills and misconceptions, a generalized diagnostic classification model for multiple-choice option-based scoring is used.

Predictor Importance in Multilevel Longitudinal Models: An Empirical Application of Dominance Analysis
Luciana Cancado, Curriculum Associates; Razia Azen, University of Wisconsin - Milwaukee

Multilevel longitudinal models may incorporate time-invariant and time-varying predictors of outcomes that change over time. After model predictors are selected, researchers might want to determine their relative importance. The application of Dominance Analysis to study the relative importance of predictors of mathematics achievement growth is demonstrated here with empirical data.

Comparing Item Parameter Drift Analysis Methods for a Computer Adaptive Test
Changjiang Wang, Pearson; David Shin, Pearson

We compare two IPD methods, one based on a narrowed distribution of student abilities, while the other based on a full distribution using pseudo counts. With this comparison, we explore whether and to what extent a narrow distribution of student abilities will impact the accurate identification of drifted items.

Equating Precision of Two Data Collection Designs: Random Groups Versus Common Items
Sooyeon Kim, Educational Testing Service; Tim Moses, The College Board

We evaluate the impact of failed randomization on equating under random-groups designs where two examinee groups take one of two forms that share common items. The groups may be nonequivalent due to unsuccessful randomization. We evaluate random-groups and common-item equating against standard errors of the differences of two equating functions.

Item Profiling: A Case Study of NAEP Grade 4 Mathematics Process Data

Glenn Hui, George Mason University; Ruhan Circi, American Institutes for Research; Soo Youn Lee, American Institutes for Research; Mingqin Zhang, The University of Iowa

NAEP is transitioning to digitally based assessments (DBAs), but some item development practices are still rooted in paper-pencil methodologies. Using NAEP DBA process data, this study investigates the possibility of identifying item profiles from process data, looking at how students interact with items that share similar characteristics.

A Class of Cognitive Diagnosis Models for Polytomous Data

Xuliang Gao, Jiangxi Normal University; Wenchao Ma, The University of Alabama - Tuscaloosa; Daxun Wang, Jiangxi Normal University; Yan Cai, Jiangxi Normal University; Dongbo Tu, Jiangxi Normal University

The current article proposes a class of CDMs for polytomous responses with less restrictive assumptions. In the proposed CDMs, three different link functions, namely, cumulative logits, local logits, and continuation ratio logits, are considered. Several commonly used polytomous CDMs can be viewed as special cases of the proposed model.

In Search of Equality: Developing an Equal Interval Likert Response Scale

Elisabeth Marie Spratto, James Madison University; Deborah L. Bandalos, James Madison University

Many response options on attitudinal scales may produce ordinal-level data rather than interval. This poses a problem for the statistical tests that may be used, as many analyses assume interval-level data. In this study, I attempted to develop a set of equal-interval Likert response options.

Comparison of Multidimensional Models for Extreme Response Styles

William Holmes Finch, Ball State University; Brian F. French, Washington State University; Maria E. Hernandez Finch, Ball State University

The study's purpose was to develop and test the performance of methods accounting for extreme response styles (ERS) in the presence of multiple latent traits. It extends earlier work that focused on ERS for unidimensional traits. Results show the new methods accurately estimate latent traits in the presence of ERS.

Benefits and Constraints of Using Out-of-Level Items in Computerized Adaptive Testing

Jie Lin, Pearson; Hua Wei, Pearson

This study investigates the extent to which item pools of different sizes (both in-level and out-of-level) can benefit from the use of out-of-level items to improve measurement accuracy and test efficiency in CAT. Results have direct implications with regards to the relevance and applicability of using out-of-level items in CAT.

Sequential or Simultaneous: Methods to Transform Numerous Chain-Linked Tests to Item Pool

Tsung-Han Ho, ETS

Simultaneous linking can be a useful approach for IRT scale linking when numerous chain-linked test forms 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.

An Investigation of the Alignment Method for Item Response Theory Scale Linking

Youhua Wei, ETS; Brendan Jackson, ETS

International testing programs often administer the same test in different countries. Some items perform differently across populations, and scale linking tends to be cumbersome for multiple populations. This study evaluates the performance of an alignment method for IRT scale linking for multiple populations without requiring exact invariant item parameters.

Handling Perfect Scores in Fixed-Length Vertically Scaled Computerized Adaptive Tests

Adam Wyse, Renaissance; Catherine Close, Renaissance Learning, Inc.; James McBride, Renaissance Learning, Inc.

A common practical challenge is how to assign ability estimates to perfect scores when using IRT models and maximum likelihood estimation. This study introduces an approach to assign ability estimates to perfect scores in fixed-length vertically-scaled computerized adaptive tests based on using Bayesian estimation methods.

Model Adequacy Checking for Applying Harmonic Regression in Assessment Quality Control Process

Jiahe Qian, ETS

When harmonic regression is applied to implementing quality control for educational assessments, it is imperative to conduct model adequacy checking for major assumptions. Three types of R-squared and two types of root mean squared errors were applied for the checking. Real data of an English-language assessment were used.

Investigating Construct Validity of the Test of Mathematical Abilities With Exceptional Students

Soyoung Park, The University of Texas - Austin

This study investigates the mean-difference scores of exceptional subgroups and a demographically matched comparison sample on the Test of Mathematical Abilities—Third Edition (TOMA-3; Brown, Cronin, and Bryant, 2013). Investigation of the differential performances of exceptional groups on a test is critical to identifying the construct validity of the test.

Investigating the Impact of Parameter Instability on Item Response Theory Proficiency Estimation

Kathleen McGrath, University of South Carolinia; Whitney Smiley, American Board of Internal Medicine; Jerome Clauser, American Board of Internal Medicine; Bradley G. Brossman, American Board of Internal Medicine

Previous research concerning proficiency estimators is based on the unrealistic assumption that items are well-calibrated. This study examines the differential performance of proficiency estimators at various levels of parameter instability to determine if different proficiency estimators are disproportionately affected by parameter instability, and if so, under what conditions this occurs.

Monitoring the Scale Stability Using Harmonic Regression

Jingyu Liu, ETS; Hanwook (Henry) Yoo, Educational Testing Service

Monitoring the scale scores across continuously administered test forms is essential to ensure the scale stability and the quality of assessment. The stability of the GRE® General Test scores are evaluated using harmonic regression approach. A residual analysis is also conducted to detect unusual score trends from outlier administrations.

Assessing Subscore Structure for Innovative Item Types Across Multiple Test Forms

Jing-Ru Xu, Pearson VUE; Joe Betts, Pearson VUE; William Joseph Muntean, Pearson

This research explores the methodology on identifying a subscore structure for innovative item types across multiple forms for a real testing program. Large-scale data with missingness were analyzed using different psychometric and statistical methods. The results exemplify how to communicate the test results efficiently to the public in test design.

Effect of Omitted or Not-Reached Missing Data Treatment on Item Response Theory Scale Linking

Zhen Li, Texas Education Agency; Haiqin Chen, American Dental Association; Mi-Suk Shim, Texas Education Agency

This study explores the impact of missing value treatments on IRT scale linking for mixed-format tests. Three classical missing value treatment methods are examined: ignore, score as incorrect, and exclude. The findings showed that the ignoring or excluding missing responses from low performers resulted in similar results.

Understanding Learner Heterogeneity: A Mixture Learning Model With Responses and Response Times

Susu Zhang, Columbia University; Shiyu Wang, University of Georgia

We propose a mixture hidden Markov Diagnostic Classifi tion Model for learning with response times and responses. It accounts for the heterogeneities in learning patterns among students by modeling the different learning and response behaviors among subgroups. The model is evaluated through a simulation study and a real data application.

State of the Profession: Historical Demographic Trends of Graduates in the Educational Measurement Field

Jennifer Randall, University of Massachusetts; Joseph Rios, University of Minnesota

This study examined the supply of educational measurement graduates by relying on historical data collected by the U.S. federal government. Specifically, we ask does the supply (earned graduate degrees) of measurement specialists represent, or reflect, the cultural, socio-cultural, and ethnic identities/histories of the constituencies they serve/assess?

The Influence of Subjective Norms on Students' Mathematical Learning Behavior

Lu Yuan, Beijing Normal University; Ying Yuan, Beijing Normal University; Xiaofeng Du, Beijing Normal University; Tao Xin, Beijing Normal University; Tuo Liu, Tianjin Normal University; Xuefeng Luo, Minnan Normal University

This research uses SEM to explore the influence of subjective norms on students' mathematical learning behavior based on the data PISA 2012 of Shanghai. The results show that the mathematical interest plays a partial mediating role between subjective norms and students' mathematical learning behavior.

Utilities of Automatic Item Generation on Parallel Form Construction

Hongwook Suh, Nebraska Department of Education; Minsung Kim, Buros Center For Testing; Jaehwa Choi, The George Washington University; Ji Hoon Ryoo, University of Southern California; Shonai Someshwar, The George Washington University

Automatic Item Generation (AIG) approach is an emerging research and practice area of generating high quality and/or massive quantity of assessment items/tests via computerized item modeling (CIM). This study investigates the utilities of AIG (i.e., parallel form reliability) in parallel form construction via Monte Carlo simulation.

National Council on Measurement in Education Invited Electronic Board 3

Andre A. Rupp, Educational Testing Service (ETS)

A Resampling Procedure for Absolute Model-Data Fit and Its Comparison With M2

Insu Paek, Florida State University; Ki Matlock Cole, Oklahoma State University; Hirotaka Fukuhara, Pearson

The performance of both parametric and nonparametric bootstrap procedures is evaluated and compared with the limited information test M2 for absolute model-data fit in detecting the misspecification of item response curve.

Sunday, April 7, 2019
5:05 – 6:35pm, Manitoba, Coordinated Session

Pioneering a New Approach to Test Design and Development

Chair: Tracy Gardner, New Meridian Corporation
Discussant: Ye Tong, Pearson

Demands from states to have more control over their test designs while taking advantage of quality content and more flexible design options are rising. As many states evolve from a consortium model to their own custom test development, vendors must develop solutions that are innovative and flexible. Coupled with the need for more flexible test designs is the push by policy makers for shorter tests and faster score reporting. In this coordinated session, the fi paper will discuss challenges facing states that have been working collectively in a national consortium. The second paper will discuss fl xible test design options that are available to states wishing to continue to use the content developed by the consortium. The third paper will discuss how adaptive testing solutions can be incorporated into the bank of items developed for the consortium. Since adaptive designs require pre-equated parameters, the fourth paper will present the challenges and solutions of a pre-equating approach and share the results of a pre-equating study. The final paper will discuss a framework called the Quality Testing Standards and Criteria for Comparability Claims (QTS), which provides guidance for states and test administration vendors that wish to make comparability claims to an established scale.

Pioneering a New Approach to Test Design and Development
Tracy Gardner, New Meridian Corporation

Alternative Blueprint Options for Test Designs With Task Models and Passage Sets
Nathan D Minchen, Pearson; Tracy Gardner, New Meridian Corporation; Aimee M. Boyd, Pearson

Integrating Computerized Adaptive Testing Models Into Test Designs With Task Models and Passage Sets
Stephen T. Murphy, Measured Progress

A Pre-Equating Study Using Test Designs With Task Models and Passage Sets
Arthur A. Thacker, Human Resources Research Organization; Erin Banjanovic, HumRRO

Ensuring Comparability to an Established Reporting Scale
Leslie Keng, National Center for the Improvement of Educational Assessment, Inc.; Erika L. Landl, Center for Assessment

Sunday, April 7, 2019
5:05 – 6:35pm, Quebec, Coordinated Session

Evaluating Teachers' Interpretation and Use of Results in Various Assessment Contexts

Chair: Priya Kannan, Educational Testing Service
Discussant: Gavin T. Brown, The University of Auckland

Throughout the school year, teachers assess their students in various ways and use the evidence of learning to inform a range of next steps. While results from summative assessments are designed to provide a snapshot of a student's learning in the previous year, evidence from formative assessment is intended to provide teachers with feedback about the gap between their students' current level of understanding and specific learning goals (Bennett, 2011; Black & William, 1998). Evidence to support the assessment's claims should be clearly articulated for the intended context and purpose within reports (Kane, 2013). Furthermore, in a utilization-oriented evaluation approach (Greene, 1988), the results presented should be evaluated based on its usefulness to the potential stakeholders. Therefore, within the overall paradigm of the interpretative/use argument (IUA), it is important to evaluate how teachers understand the information presented in score reports and use it to inform their instructional decisions and practice. In light of this year's conference theme, the collection of papers in this session will focus on communicating assessment results effectively to teachers in various assessment contexts by focusing on the ways in which teachers interpret and use the results provided in each of these contexts to inform their instruction.

Teacher-Centric Design Process for a Dashboard to Support Formative Assessment
Robert Dolan, CAST; Cara Wojcik, CAST; Emma Starr, CAST; Kim Ducharme, CAST, Inc.; Jose Blackorby, CAST

Helping Teachers Make Sense of Data in Formative Contexts
E. Caroline Wylie, ETS; Christine Jennifer Lyon, ETS

Teachers' Interpretations of Interim Assessment Results Presented in a Learning Progressions Framework
Priya Kannan, Educational Testing Service; Andrew Bryant, ETS; Vera Shao, Educational Testing Service; E. Caroline Wylie, ETS

Understanding Teachers' Interpretation of Achievement Labels From Score Reports
Francis O'Donnell, University of Massachusetts, Amherst; Stephen G. Sireci, University of Massachusetts - Amherst; April L. Zenisky, University of Massachusetts - Amherst

Sunday, April 7, 2019
5:05 – 6:35pm, Salon A, Paper Session

New Insights on Engagement, Learning, and Performance

Discussant: Mike K. Russell, Boston College

Mediation Effect of Self-Efficacy on the Predictive Validity of TOEFL iBT Scores

Ya Zhang, Western Michigan University; Suhayb Kattan, Western Michigan University; Wessam Abdelaziz, Western Michigan University

A series of structure equation models were used to examine the mediation effect of self-efficacy and academic stress on the predictive validity of TOEFL iBT scores. The mediation effects were evaluated across time and academic disciplines. The study provided insights into the relationship between TOEFL scores and academic success.

Effect of Class Engagement on Academic Achievement Through Assessment-Oriented Learning

Sun Geun Baek, Seoul National University; Yun-Kyung Kim, Seoul National University

The effect of class engagement (CE) on academic achievement (AA) through assessment-oriented learning (AOL) was investigated with the data of 434 high school students. As a result, the effect of CE on AA was fully mediated by AOL, implying that AOL should be considered a crucial factor to improve AA.

Is Interactivity Enough or Do We Need Game Elements to Enhance Engagement?

Sue Ward, ACT; Meirav Arieli-Attali, ACT, Inc.

We will present a study that incorporates interactivity with and without game elements to a science assessment. We investigate the effect of various features on student performance and engagement, via a comparison of four variations of the same assessment. We will report results from a study with middle school students.

New Methods to Detect Low-Motivation Responses on Low-Stakes Tests

Xiao Luo, Measured Progress; Xi Wang, Measured Progress; Louis A. Roussos, Measured Progress

A scientifically sound method for detecting low-motivation responses is long desired, given its prevalence in low-stakes tests. This study proposes three low-motivation detection methods and examines their efficacy against a real data. The proposed methods are expected to help produce unbiased results, draw more valid inferences, and improve test validity.

Sunday, April 7, 2019
5:05 – 6:35pm, Salon B, Invited Speaker Session

Using the ACT and SAT for Accountability Under the Every Student Succeeds Act: Appropriate or Inappropriate Use

Panelist: Wayne J. Camara, ACT, Inc.
Panelist: Scott F. Marion, National Center for the Improvement of Educational Assessment, Inc.
Panelist: Denny Way, College Board
Panelist: Dale Whittington, Retired
Moderator: Sean "Jack" P. Buckley, American Institutes for Research

ESEA calls for rigorous college-and career-ready standards and assessments, but also provides districts/LEAs additional flexibility to request state approval to use a nationally recognized assessment in lieu of the state high school assessment. ESEA specifically acknowledges its intent is to recognize assessments, such as the ACT and SAT, as desirable alternatives, subject to approval. This was partially designed to reduce overall testing burden. However, USED has established requirements as part of the peer review process. Specifically, the requirements call for the national assessment to be equivalent or more rigorous with respect to: (a) content coverage of academic standards, (b) difficulty, (c) cognitive complexity, (d) overall quality, and (e) validity and reliability. At this time, a number of interpretive issues remain unresolved concerning the requirements and how they will be evaluated under peer review when states select a national admissions test. Coverage of content standards, or alignment are often considered the most difficult hurdle because college admissions tests are relatively much shorter with fewer items than state assessments and have a single performance task. Today, nearly half of all states offer admissions tests to all public school students, and about half of these states plan to use scores for federal accountability (Gewertz, 2017). This debate will focus on the pros and cons of such uses for the ACT and SAT.

Sunday, April 7, 2019
5:05 – 6:35pm, Territories, Paper Session

Technical Considerations in Measuring Social and Emotional Learning

Discussant: Patrick Charles Kyllonen, Educational Testing Service

Psychometric Evaluation of Social Emotional Learning Measures: Applying Multidimensional Item Response Theory

Youngsoon Kang, University of Minnesota - Twin Cities; Michael C. Rodriguez, University of Minnesota; Kory Vue, University of Minnesota - Twin Cities

Measures of social and emotional learning (SEL) are gaining interest nationally. However, they have not been rigorously evaluated psychometrically. As part of a larger effort to develop psychometric evaluation criteria, we explore a large-scale measure of SEL through multiple models of confirmatory factor analysis and item response theory.

Measurement Invariance of an International Developmental Assets Measure: Alignment of 29 Countries

Rik Lamm, University of Minnesota - Twin Cities; Tai Tri Do, University of Minnesota - Twin Cities; Michael C. Rodriguez, University of Minnesota; Peter C. Scales; Eugene Roehlkepartain, Search Institute

The Developmental Asset Profile is used worldwide to measure social and emotional learning. No validity evidence exists regarding measurement invariance across countries. Multigroup Alignment was used with data from 29 countries to identify items and countries that showed invariance, ultimately to estimate and compare country means.

Validation of Social and Emotional Learning Measures in Inequitable Settings

Michael C. Rodriguez, University of Minnesota; Michael Dosedel, University of Minnesota - Twin Cities; Youngsoon Kang, University of Minnesota - Twin Cities

In school contexts, validation of social and emotional learning measures should be relevant to important school outcomes (via an interpretation/use argument). In systems with significant inequities in school outcomes (e.g., discipline and suspension rates), there is significant differential prediction. However, this is explained by differences in incidence rates by group.

Evaluating the Differentiation of Social-Emotional Learning Constructs Using Multilevel Factor Analysis

Daniel M. Bolt, University of Wisconsin - Madison; Caroline Wang; Robert H. Meyer, Education Analytics

We applied multilevel factor analysis to self-report rating scale data from over 400,000 students and 1600 schools to examine differentiation among four SEL constructs at school and student levels from 3rd-12th grade. The measures are further studied in their prediction of academic achievement as well as measurement invariance across levels.

Understanding Response Processes in Noncognitive Measures With Explanatory Item Response Modeling

Okan Bulut, University of Alberta; Hilal Celik, Marmara University; Ming Lei, American Institutes for Research

A common criticism of noncognitive measures is the potential for cultural influences in response processes. This study employs explanatory item response modeling to understand the effects of various demographic variables on response processes, using a measure of social support. Results support validity concerns due to high sensitivity of noncognitive measures.

Monday, April 8, 2019
8:00 – 10:00am, Alberta, Invited Speaker Session

Appropriately Interpreting, Comparing, and Communicating Results From International Assessments: Challenges and Opportunities

Chair: Leslie Rutkowski, Indiana University
Chair: Matthias Von Davier, National Board of Medical Examiners

In the past twenty-odd years, international educational assessments have grown in terms of number of studies, cycles, and participating countries, many of which are a heterogeneous mix of economies, languages, cultures, and geography. For example, PISA has grown from 43 participating educational systems in 2000 to 72 participating populations in 2015. In a similar vein, content and platforms have evolved in many international assessments over the past two decades. From paper-and-pencil tests to multistage designs to an emphasis on innovative domains such as collaborative problem solving in 2015 and global competence in 2018, much has changed in the international assessment landscape. Taken together with the fact that these snapshots of achievement are cross-sectional measurements of some target population on one day in limited content domains, appropriately interpreting performance within and across educational systems is a challenge. To that end, we bring together recognized scholars in the field of educational measurement, economics, and educational policy to discuss the inherent challenges of communicating results from international assessments in the 21st century.

Fallacies in Interpretation? How Simple Achievement Rankings Can Mislead Efforts to Improve Educational Outcomes
Kadriye Ercikan, Educational Testing Service, Princeton, NJ 08541

What the Data Can Support: Avoiding Pitfalls in Interpreting International Assessment Results
Henry I. Braun, Boston College

Are International Assessment Scores the Key to Worker Productivity Among Industrialized Nations?
Henry M. Levin, Teachers College, Columbia University

Disaggregating International Assessment Results: Are America's Schools Really Failing?
David C. Berliner, Arizona State University

Monday, April 8, 2019
8:00–10:00am, Algonquin, Coordinated Session

Optimizing Digital Affordances in NAEP Assessment Tasks: Findings From Two Research Studies

Chair: Fran Stancavage, American Institutes for Research
Discussant: Jesse R. Sparks, Educational Testing Service

In this session we present fi from two studies that examine the impact of digital affordances in NAEP extended scenario-based tasks (SBTs) in science and reading. The first is a cognitive lab study that investigates which key visual and associated interactive features of NAEP Science SBTs might inhibit or enable the ability of students to accurately demonstrate their actual level of mastery of target knowledge and skills. The second is a randomized control trial that focuses on the overall impact of NAEP Reading SBTs on students' reading performance, reading behaviors, and engagement by comparing outcomes for Reading SBTs with outcomes for traditional NAEP discrete (DI) blocks using the same texts and items as the SBTs, but without any of the SBT features (e.g., purpose setting, avatars). The opening presentation provides context for the research studies by describing the ways in which NAEP is integrating digital affordances into items in every subject area. The concluding presentation speaks to next steps by proposing the development of a comprehensive research agenda to systematically examine the impact of a broad range of multimedia and interactive features in assessment.

Overview of Efforts to Integrate Digital Affordances Into NAEP Item Development
Peggy G. Carr, National Center for Education Statistics/IES, U.S. Department of Education

Design of the Cognitive Lab Study of NAEP Science Scenario-Based Tasks
Richard P. Duran, University of California - Santa Barbara; Ting Zhang, American Institutes for Research

Findings From the Cognitive Lab Study of NAEP Science Scenario-Based Tasks
Ting Zhang, American Institutes for Research; David Sanosa, University of California - Santa Barbara

Key Findings From the Randomized Control Trial of NAEP Reading Scenario-Based Tasks
Karen Wixson, ETS; Hilary Persky, Educational Testing Service; David S. Freund, ETS

Analyzing Process Data for the Randomized Control Trial of NAEP Reading Scenario-Based Tasks
Gary Feng, Educational Testing Service; Susan Shuai, ETS; Chris Agard, Educational Testing Service

Development of a Comprehensive Research Agenda to Examine the Impact of Digital Affordances in Assessment
Fran Stancavage, American Institutes for Research

Monday, April 8, 2019
8:00 – 10:00am, Ballroom, Coordinated Session

Raising a New Generation of Measurement Experts: Stories From Around the World

Chair: Tzur M. Karelitz, National Institute for Testing & Evaluation
Discussant: Avi Allalouf, National Institute for Testing and Evaluation (NITE)

Across the world, there is a growing need for psychometricians – experts in educational and psychological measurement. Psychometrics is a relatively obscure profession. Most people are unaware of, or uninterested in, psychometrics. Indeed, most countries (apart from the US and a few other countries) lack academic programs or professional training in this field. As the role of testing becomes more prominent in today's society, the shortage of measurement experts becomes a serious problem. This symposium will showcase some countries (Sweden, Israel, Spain, Russia, Australia, Norway, and the US) that are dealing with the shortage of measurement experts. Participants will discuss what lead to the shortage of experts in their country, how they attempted to address the shortage and what challenges they faced in the process. We believe that presenting this topic to the measurement community might enhance and expedite international efforts to develop and promote the field of psychometrics in other countries.

Raising a New Generation of Measurement Experts: The Case of Sweden
Christina Wikstrom, Umea University

Raising a New Generation of Measurement Experts: The Case of Israel
Tzur M. Karelitz, National Institute for Testing & Evaluation; Avi Allalouf, National Institute for Testing and Evaluation (NITE)

Raising a New Generation of Measurement Experts: The Case of Spain
José Muñiz, University of Oviedo

Raising a New Generation of Measurement Experts: The Case of Russia
Alina Ivanova, National Research University Higher School of Economics; Elena Kardanova, National Research University

Raising a New Generation of Measurement Experts: The Case of Australia
Goran Lazendic, Australian Curriculum, Assessment and Reporting Authority

Raising a New Generation of Measurement Experts: The Case of Norway
Rolf Olsen, CEMO, University of Oslo

Raising a New Generation of Measurement Experts: The Case of the United States
Stephen G. Sireci, University of Massachusetts - Amherst; Yooyoung Park, University of Massachusetts - Amherst

Monday, April 8, 2019
8:00 – 10:00am, British Columbia, Paper Session

Advances in Test Security

Discussant: John Fremer, Caveon Test Security

Mining Process Data to Detect Item Harvesters

Manqian Liao, University of Maryland - College Park; Jeffrey Patton, Financial Industry Regulatory Authority; Ray Y Yan, Financial Industry Regulatory Authority; Hong Jiao, University of Maryland-College Park

Item harvesting behaviors are difficult to detect by existing statistical approaches due to the idiosyncratic nature of human behaviors and the absence of operational definitions. This study develops a data mining approach to discover behavioral archetypes and detect abnormal patterns, which could be used as preliminary flags for item harvesters.

Combining Marginal and Conditional Exposure Control in Adaptive Testing

Qi Diao, ETS; Hao Ren, Pearson

The dilemma of conditional exposure control methods is that in real world, the true ability of the test taker is unknown while the methods are trying to control the exposure of test takes of similar ability levels. The proposed method combines marginal exposure control with conditional exposure control.

Study of Graph Theory Approach to Detect Pre-Knowledge Using Real, Marked Data

Dmitry Belov, Law School Admission Council; Sarah Linnea Toton, Caveon Test Security

An experiment was conducted to embed item pre-knowledge into a group of examinees, where most examinees simply took a test, but some examinees had access to a subset of items before the exam. The resulting real dataset was used to study a recently developed graph theory based detector of pre-knowledge.

Detection of Compromised Items in a Computerized Adaptive Testing Licensure Exam Using Sequential Procedures

Chansoon (Danielle) Lee, National Council of State Boards of Nursing; Qian Hong, NCSBN

This study applies real-time monitoring sequential procedures to a real operational item pool in CAT to detect compromised items. The sequential procedures examine changes in the individual item response function using a series of statistical hypothesis tests based on CTT and IRT.

Test and Item Time Metrics for Test Security Evaluation

Kirk A. Becker, Pearson; Qing Yi, Pearson

Computer based testing records time spent on items. Security research suggests that timing analysis can be used to detect cheating. This study will look at timing data, both general samples and data where cheating was detected. We will provide baseline data on timing statistics and evaluate a new timing analysis.

Monday, April 8, 2019
8:00 – 10:00am, Manitoba, Paper Session

Advances in Evaluating Psychometric Models

Discussant: Seock-Ho Kim, University of Georgia

Two New Item-Fit Statistics for the Lognormal Model for Response Times

Sandip Sinharay, Educational Testing Service; Peter Van Rijn, ETS Global

Two new item-fit statistics are suggested for the lognormal model for response times (van der Linden, 2006). The theoretical large-sample distributions of the statistics under no mis fi are derived. The properties of the new statistics are examined using simulated and real data.

Level-Specific Evaluation of Model Fit in Item Response Theory

Scott Monroe, University of Massachusetts - Amherst; Megan Kuhfeld, NWEA; Nermin Kibrislioglu Uysal, Hacettepe University

This study proposes level-specifi evaluation of model fi for IRT models using a recently-developed variant of posterior predictive model checking which assumes normality of the posterior distribution. Test quantities motivated by limited-information goodness-of-fit testing are used with clustered data. The method is demonstrated via a simulation study.

Impact of Latent Regression Model Complexities on Group Score Estimates

Nuo Xi, Educational Testing Service; John R. Donoghue, Educational Testing Service; Yue Jia, Educational Testing Service

The research objective is to examine the impact on group score estimates when the complexity of the fitted latent regression model is a control factor. Simulation studies are designed and conducted to evaluate this impact and provide feedback on the latent regression model setup applied in the current NAEP operations.

Anchoring Rater Effects From a Suboptimal Judging Plan: A Sensitivity Analysis

Christopher T. Moore, Minneapolis Public Schools

Rater-mediated assessment programs strive to implement judging plans that adequately connect each rater to others. Disconnectivity can be addressed by combining ratings from multiple time periods. To what degree are many-facet Rasch scores of teaching ability sensitive to calibration with pooled ratings? What are the consequences of ignoring rater variation?

Monday, April 8, 2019
8:00 – 10:00am, Quebec, Paper Session

Applied Issues in Large-Scale Assessments

Discussant: Michael C. Rodriguez, University of Minnesota

Reducing Testing Time: One State's Approach to Revising the Blueprint

Joseph Fitzpatrick, NCS Pearson, Inc.; Joyce Zurkowski, Colorado Department of Education; Jennifer Beimers, Pearson; Jasmine Carey, Colorado Department of Education

In 2018, Colorado implemented revised assessments based on reduced versions of the PARCC blueprints. This paper describes several analyses conducted to help determine the measurement impact of this revision. Much of the focus is on how these results were placed in the context of the goals of the assessment program.

Consequences of Model Misfit on Reporting Outcomes for a Large-Scale Assessment

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

This study examines consequences of model misfit on reporting outcomes for a large scale Algebra 1 End of Course assessment. It investigates whether ignoring misfit at the item level has any practical impact on scale scores, their standard errors, and performance level classifications reported for parents, classroom teachers, and educators.

Using Confidence-Based Testing in Large-Scale Assessments

Yiran Chen, University of Michigan - Ann Arbor

Using simulated data, this paper evaluates the potential of confidence-based testing (CBT). The results suggest that CBT substantially improves test reliability, while leaves population-level estimates unbiased. The fi suggest that CBT may allow large-scale assessment programs to safely halve the required sample size, saving millions of dollars in field operation.

Determining Optimal Bounds of Item Response Theory Scores for Developing a Meaningful Score Scale

Edison M. Choe, The Graduate Management Admission Council; Kyung (Chris) T. Han, The Graduate Management Admission Council

In operational testing programs, IRT scores (theta) are most commonly estimated using maximum likelihood. This requires defining the parameter space, which is often done arbitrarily by imposing uninformed bounds that are symmetric about zero. We investigate several methods that systematically determine optimal theta bounds for developing a meaningful score scale.

Transition to Item Response Theory for an Existing CTT-Based Assessment

Weiwei Cui, The College Board; YoungKoung Kim, The College Board; Tim Moses, The College Board

This study evaluates the above two psychometric properties of two methods to link the IRT ability estimates to existing reporting scale scores and the two procedures for preserving the content alignment across assessments, especially focusing on the impact of the prior distribution of latent abilities.

Monday, April 8, 2019
8:00 – 10:00am, Salon A, Coordinated Session

Best Practices Around Automated Scoring Standards

Chair: Andre A. Rupp, Educational Testing Service (ETS)
Discussant: Andre A. Rupp, Educational Testing Service (ETS)

Without a doubt, the use of automated scoring technology writ large is becoming more and more prevalent in today's testing programs. Many organizations are utilizing automated technologies to quickly identify and synthesize construct-related evidence, to produce scores, and to create diagnostic feedback. As this technology evolves and becomes more frequently used in large-scale assessment and learning environments, it becomes crucial that scientifi defensible (i.e., robust) and practically meaningful (i.e., actionable) standards for its development and implementation are co-created in the constituent communities who work on these systems. Importantly, standards need to be instantiated through best practices within institutions who aim to adhere to the standards, which, in turn, involves the creation of workflows, team compositions, artifacts, and mindsets that reflect the mission of these standards. Yet it is rare to be afforded a cross-institutional "look behind the scenes" of such instantiations to help others learn from what works and what does not work in these context. This session is designed to unveil some of these hidden secrets and bring together experts from diverse institutions with diverse assessment and learning programs and diverse practices. The format is different from other coordinated sessions as it involves a mixture of four sequential presentations, four parallel roundtables, and a moderated panel discussion to maximize audience engagement.

Best Practices for Automated Scoring at ACT and ACTNext
Erin Yao, ACT; Scott William Wood, ACT, Inc.; Pravin V Chopade, ACT, Inc.; Saad Khan, ACTNext

Best Practices for Automated Scoring at AIR
Sue Lottridge, American Institutes of Research

Best Practices for Automated Scoring at ETS
Cathy LW Wendler, Educational Testing Service; Andre A. Rupp, Educational Testing Service (ETS)

Best Practices for Automated Scoring at Pearson
Peter W. Foltz, Pearson; Kyle Habermehl, Pearson Education, Inc.

Monday, April 8, 2019
8:00 – 10:00am, Salon B, Paper Session

Technical Considerations in Calculating and Evaluating Reliability

Discussant: Michael E. Walker, The College Board

Examining Rating Designs With Cross-Classification Multilevel Rasch Models

Jue Wang, University of Miami; Zhenqiu Lu, University of Georgia; George Engelhard, University of Georgia; Allan S. Cohen, University of Georgia

A simulation study is designed and implemented to examine the effect of different rating designs on parameter estimates for the Cross-Classification Multilevel Rasch Model. Results indicated that the incompleteness of a design affected the accuracy and effectiveness of the model, but a larger sample size can greatly improve rater estimates.

Can Less Be More? The Relationship Between Test Length and Reliability

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

We investigated the performance of a pretesting method that retained only the best-performing items, rather than removing only defective items. Using this method, we examined the relationship between test length and reliability. For some conditions, we achieved comparable or better reliability using fewer live items (i.e., more pretest items).

Calculating Conditional Reliability for Dynamic Measurement Model Capacity Estimates

Denis Dumas, University of Denver; Daniel McNeish, Arizona State University

Dynamic Measurement Modeling (DMM) is a recent framework for measuring developing constructs whose manifestation occurs after an assessment is administered (e.g., learning capacity). This paper advances one method for computing conditional reliability for DMM capacity scores so that precision of the estimates can be assessed, and tests the efficacy of that reliability method via a simulation study.

Quantifying Reliability for Oral Reading Fluency Assessment Using the Grubbs Model

Cornelis Potgieter, Southern Methodist University; Akihito Kamata, Southern Methodist University

This study proposes to apply the Grubbs model to estimate oral reading fluency more accurately. It is demonstrated that application of the Grubbs model allows quantifying measurement error variance to determine optimal weights to combine passage-level data to improve the fluency score for each student.

A Transdisciplinary View of Measurement Error Models and the Variations of X = T + E

Bruno D. Zumbo, The University of British Columbia; Edward Kroc, University of British Columbia

The purpose of this paper is to describe the connection between five linearly additive measurement error models. With any eye to informing practitioners, we show that although these models are deceptively similar in their general algebraic form, X=T+ E, they have different error structures that both connect and distinguish them.

Monday, April 8, 2019
8:00 – 10:00am, Territories, Paper Session

Examining Impacts of Rater Effects

Discussant: Melinda A Taylor, ACT, Inc.

Evaluating the Impact of Rater Effects on Item Response Theory–Based Scoring: It Matters

Andrea Gotzmann, Medical Council of Canada; Sirius Qin, Medical Council of Canada; Maxim Morin, Medical Council of Canada; André F. De Champlain, Medical Council of Canada

Evaluating when and how to estimate rater effects within FACETS is complex. Ten rater effects, two score types, and three rater sizes were simulated and scored using three FACETS models. This study shows performance assessments should incorporate rater estimates and carefully consider which FACET scoring model to apply.

Combined Effects of Rater Misfit and Differential Rater Functioning in Performance Assessments

Wenjing Guo, The University of Alabama; Stefanie Wind, The University of Alabama

We explored combinations of rater misfit and differential rater functioning (DRF) in performance assessment. We considered the sensitivity of misfit and DRF indices when both effects were present, and in sparse designs. Analyses revealed challenges in disentangling the effects using only numeric indicators. Residual analyses provided insight into rater idiosyncrasies.

The Impact of Multiple Rater Effects on Different Methods of Adjusting Scores

Thai Ong, James Madison University; Jason P. Kopp, American Board of Surgery; Andrew Jones, American Board of Surgery

We evaluated the impact of four rater effects and their interactions on the classification accuracy of unadjusted scores, adjusted scores based on the Rasch method, and adjusted scores based on the Deviation method. Results indicated the Deviation method outperformed the unadjusted and Rasch methods in most simulated conditions.

Personalized Feedback Influences Rating Accuracy in Online Essay Scoring

Ma Jie, Beijing Normal University; Hongyun Liu, Beijing Normal University

This study investigated the current situation of online essay scoring and how personalized feedback influenced rating accuracy, severity and consistency. Results revealed compared with expert, original raters are more lenient. Compared with non-feedback group, CTT feedback can hold the standard deviation while MRFM feedback can reduce the mean rating score.

Understanding and Mitigating Rater Drift: When Does Rater Drift Threaten Score Comparability?

Michelle Boyer, Data Recognition Corporation; Richard J. Patz, University of California Berkeley

Rater accuracy and consistency are fundamental to test score validity, yet rater behavior is routinely evaluated with relatively ad hoc procedures. A systematic and coherent model for the types of errors that raters might make is used to facilitate a deeper understanding and mitigation of rater drift in test equating.

Monday, April 8, 2019
10:25 – 11:55am, Alberta, Invited Speaker Session

Equity-Centered Design in Assessment: Diversity Issues in Testing Committee Invited Session

Chair: Jennifer Randall, University of Massachusetts
Panelist: Antonia Darder, Loyola Marymount University
Panelist: Ezekiel J. Dixon-Roman, The University of Pennsylvania
Panelist: Jamila Lyiscott, University of Massachusetts – Amherst
Panelist: Maria Elena Oliveri, Educational Testing Service
Moderator: Joseph Rios, University of Minnesota

In this Invited Session we challenge panelists and the audience to imagine an assessment system in which students of color are centered. Historically, the assessment community has relied primarily on bias panels (post-item development) and statistical indices (post-test administration) to identify possible sources of differential performance patterns. In this session, we invite panelists and the audience to ponder how we can create an assessment system culture in which the sociocultural identities of students are deliberately considered and valued– not as an afterthought, but rather- in the planning and development phases of assessment. The ultimate goal of this session is for policy makers and test developers to begin to think more critically about the ways in which we assess students of color and, consequently, move towards a more fair and equitable system for these students.

Monday, April 8, 2019
10:25–11:55am, Algonquin, Coordinated Session

Evidence-Centered Design Extensions: Research in EdTech and Game-Based Learning and Assessment

Chair: Ada Woo, ACT Inc
Discussant: Alina A. Von Davier, ACT, Inc.

In the past few decades, Evidence-Centered Design (ECD) has been adopted more widely by large scale assessment organizations as a conceptual framework for test development. In the ECD approach, assessments are constructed based on evidentiary arguments. ECD provides a structured approach to develop assessments that address specific claims demonstrating that learning on a specific domain has taken place. The strength of this evidentiary approach fits well with the development of learning and technology-enhanced assessment instruments. The ECD approach is being extended to the development of learning and assessment systems (LAS) as well as other education technology (edtech) research. In the current session, researchers from four measurement and edtech organizations will share examples on how the ECD approach is being extended. Speakers will discuss ECD applications in the development of game-based LAS, adaptive LAS, and Learner Centered Design (LCD) for LAS, as well as present examples of these LAS of the 21st Century.

Evidence-Centered Design for Games
Kristen E. Dicerbo, Pearson

Extending Evidence-Centered Design for Quality Game-Based Assessments for Learning
Seth Corrigan, LRNG-GlassLab

Developing a Learning and Assessment System for Learning With the Expanded ECD Framework
Meirav Arieli-Attali, ACTNext by ACT, Inc.

Learner-Centered Design: Developing Student-Centered Assessments
Amanda Newlin, Smart Sparrow; Heather Newlin, Smart Sparrow

Monday, April 8, 2019
10:25 – 11:55am, Ballroom, Coordinated Session

Facilitating the National Research Council's Assessment Triangle

Chair: Mark R. Wilson, University of California - Berkeley
Discussant: Michelle LaMar, Educational Testing Service

In its foundational volume, Knowing What Students Know (NRC, 2001), the NRC posited a conception of how educational assessment should function, and used its Assessment Triangle as a representation of that idea. In this symposium, we describe a software system based on that conceptualization, and discuss issues that arise in that context. The BEAR Assessment System Software (BASS) encompasses a theoretical basis for the domain modeling logic, and offers tools for development, delivery, scoring, reporting and use of learning evidence. BASS employs the UC Berkeley BEAR Assessment System (BAS) to coordinate the instrumentation with the domain modeling, which is a four-part approach to modeling that involves a combined domain and student model. The structure of the software system is designed to highlight the four parts of the BAS, and to allow educational practitioners and developers to implement the processes. In this symposium, we first present and demo BASS 2.0. Then we elaborate on its approach to construct mapping of the domain and student models, and how these serve as a basis for the "full cycle" of instrument development, deployment and reporting. Then we discuss how the software enables educational practitioners and developers to implement the full range of assessment processes.

The BEAR Assessment System Software 2.0 (BASS)
David Torres-Irribarra, MIDE, Pontificia Universidad Católica de Chile; Mark R. Wilson, University of California - Berkeley; Karen L. Draney, University of California - Berkeley

Challenges of Full-Cycle Assessment Facilitation
Mark R. Wilson, University of California - Berkeley; David Torres-Irribarra, MIDE, Pontificia Universidad Católica de Chile; Karen L. Draney, University of California - Berkeley

The End User Perspective on Full-Cycle Assessment Facilitation
Rebecca Freund, University of California - Berkeley; Amy Elizabeth Arneson, University of California, Berkeley; Mark R. Wilson, University of California - Berkeley

Monday, April 8, 2019
10:25 – 11:55am, British Columbia, Invited Speaker Session

Preparing Students for College and Careers: Theory, Measurement, and Educational Practice

Discussant: Matthew Gaertner
Chair: Krista D. Mattern, ACT, Inc.
Chair: Katie Larsen McClarty, Questar Assessment Inc.

This session highlights five chapters from a recently published book in the NCME Book Series, Preparing Students for College and Careers: Theory, Measurement, and Educational Practice. The book synthesizes the current state of college- and career-readiness research and best practice in measurement and highlights how scientific rigor from psychometrics, measurement, and educational research can come together to support college and career readiness for all students. In this session, the first paper argues for a shift – from defining and measuring college readiness as a set of discrete knowledge and skills to a set of enduring practices and ways of thinking. The second paper details an empirical process for setting college- and career-readiness performance standards and for linking those standards down through lower grade levels. The third paper tackles the issue of fairness in college- and career-readiness assessments. The fourth paper describes social psychological interventions that are effective at improving college- and career-readiness. The fifth illustrates how changes in formative classroom assessment can enable students to take more control of their learning outcomes and better prepare for college and careers.

Conceptualizing and Measuring Progress Toward College and Career Readiness in Mathematics
James W. Pellegrino, University of Illinois at Chicago; William G. McCallum, The University of Arizona

Establishing Benchmarks, Cut Scores, and Performance-Level Descriptors on the Basis of Empirical Data
Wayne J. Camara, ACT, Inc.; Jeff Michael Allen, ACT, Inc.; Joann Moore, ACT

Fairness Issues in the Assessment of College and Career Readiness
Rebecca Zwick, Educational Testing Service

Supporting College and Career Readiness Through Social Psychological Interventions
Kathryn M Kroeper, Indiana University; Mary C. Murphy, Indiana University

Changing the Assessment Relationship to Empower Teachers and Students
Margaret Heritage, WestEd

Monday, April 8, 2019
10:25 – 11:55am, Imperial Room, Graduate Student Poster Session

GSIC Graduate Student Poster Session 2

Approaches to Analyzing Innovative Language Assessment Items.
David Alpizar, Washington State University - Pullman; Tongyun Li, Educational Testing Service

C-test is an innovative item type that has the potential to measure second language proficiency more efficiently than traditional multiple choice items. However, literature is scarce on examining the psychometric properties of C-test items. This study examines the local dependency issue and the best scoring model for the C-test items.

Comparing Combinations of Variable Standardization Methods and Clustering Algorithms With Real Data
Xiaoliang Zhou, Teachers College, Columbia University

I used the Iris data to compare variable standardizations and clustering algorithms. I found that standardizing by sum and group average algorithm generally produced the best ARI, algorithms were robust against some standardization methods, and the selection of potentially group-separable variables improved fitting for most algorithms, particularly the Ward's method.

The Impact of Extreme Response Style on the Result of Mean Comparison
Zhaoxi Yang, Beijing Normal University; Yingbin Zhang, University of Illinois at Urbana - Champaign; Yehui Wang, Beijing Normal University

Extreme response style (ERS) is prevalent in survey research using Likert scales. The simulation study showed that ERS could lead to a biased estimate of the difference between two groups in the interest variable, especially when ERS differed greatly between groups. An empirical example replicated the result.

Computer Adaptive Test Incorporating Incorrect Option Information
Wei Schneider, The University of Iowa

This study focuses on utilizing information provided by incorrect multiple choice item options in computer adaptive testing. Different incorrect options appeal to examinees at different levels of ability, and this information may be used to improve item selection, which in turn may signifi tly improve measurement accuracy and reduce test length.

Social Consequences: A Review From 1989 to Current
Darius D Taylor, University of Massachusetts - Amherst

This paper surveys the literature to assess the prevalence of past and recent studies (beyond DIF) that evaluate the social consequences of test score use since the release of Messick's (1989) validity chapter. Based on this review we recommend the measurement community bolster current test validation efforts.

Content Validity of a Mathematics Placement Test for a Gifted High School
Hannah Ruth Anderson, Illinois Mathematics and Science Academy

The Content Validity of a mathematics placement test at a Science, Technology, Engineering, and Mathematics (STEM) gifted residential high school is examined. Data were collected from internal and external mathematics subject matter experts (SMEs) using a card-sorting task and were analyzed using Multidimensional Scaling (MDS) and Hierarchical Cluster Analysis (HCA).

A Taxonomy for Assessing 21st Century Skills in Learning

Dandan Chen, University of Connecticut - Storrs

This study argues with supporting evidence that instituting such a taxonomy is indispensable to meeting the urging needs of 21st century skill assessment. Also, it brings up a unique taxonomy, the Taxonomy for Assessing 21st Century Skills (TA21CS).

Detection of Item Parameter Drift and Item Exposure Using Change Point Analysis

Xiaodan Tang, University of Illinois at Chicago; Haiqin Chen

Change point analysis (CPA) aims at detecting aberrant changes during a time series process. This study simulated the context of item exposure and proposed to use CPA to detect aberrant item parameter drift during multiple administrations of the same CAT item bank. We considered different conditions of item exposure rate.

Comparing Two Methods for Detecting Mode Effect Between Paper-Based and Computer-Based Assessments

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

This study aims to compare two methods (one-step method and two-step method) for detecting mode effect between paper- and computer-based assessments using data from PISA 2015 field study. The results can help measurement researchers and practitioners gain a deeper understanding of the differences in performance caused by different test modes.

Model Selection and Classification Accuracy in Cognitive Diagnosis

Yanhong Bian, Rutgers, the State University of New Jersey; Chia-Yi Chiu, Rutgers University

This study explores the performance of six literally prevalent model-fit indices or tests in cognitive diagnosis and how is the classification accuracy of the selected model compared with the general models and nonparametric methods. A simulation study was conducted under a variety of conditions and results were presented.

Differential Measurement of Science Teaching Efficacy and Beliefs: A Polytomous Differential Item Functioning Study

Neda Moslemi, University of Saskatchewan; Laurie-Ann M. Hellsten, University of Saskatchewan

Measurement equivalence is one of the most important aspects of fairness in testing, and differential item functioning (DIF) is one technique used to assess it. We evaluated DIF in the STEBI self-reported teaching efficacy instrument using Ordinal Logistic Regression and found several items exhibiting DIF of considerable magnitude.

Classification Consistency and Accuracy of Student Growth Percentiles Using Item Response Theory Methods

Adam Reeger, University of Iowa; Ariel M. Aloe, University of Iowa

This study investigates classification consistency and accuracy for student growth percentiles (SGPs) as a function of test length and ability distribution. As more states mandate measures of growth be included with proficiency, it is important to understand the implications of classifying students into categories based on growth cutscores.

Using a Support Vector Machine to Categorize Automatically Generated Items

Stephanie Varga, University of Alberta; Mark Gierl

Quality test items require significant resources in time, effort, and money to create. We propose using a support vector machine to assist with the content classification of automatically generated items. Preliminary results show a high accuracy of classification. In practice, this will improve the test development process.

Differential Response Processes Among Adults From Different Age Groups on the Program for the International Assessment of Adult Competencies

Jacquelyn A Bialo, Georgia State University; Hongli Li, Georgia State University; Jingxuan Liu, Georgia State University

Adults tend to process information more slowly and less efficiently as they age. In this study, we use response process data from the Programme for the International Assessment of Adult Competencies (PIAAC) to examine if younger, middle-aged, and older adults demonstrate differential response processes.

Vertical Equating With Longitudinal Data: Unidimensional and Multidimensional Item Response Theory Models

Yan Yan, Georgia Tech; Susan Embretson, Georgia Institute of Technology

This study examines the performance of several unidimensional and multidimensional item response theory models on the vertical linking of 7th and 8th grades mathematics tests with longitudinal design. It was found a multidimensional item response theory model for change yielded the most plausible results for equating.

How to Incorporate Response Times in Automated Test Assembly

Benjamin Becker, Institute for Educational Quality Improvement (IQB); Sebastian Weirich; Dries Debeer, University of Zurich; Frank Goldhammer, DIPF | Leibniz Institute for Research and Information in Education

In high-takes assessments test forms have to be balanced regarding test length. We propose to use a generalization of the Hierarchical Response Time Model in the Automated Test Assembly, which additionally estimates an item discrimination parameter. Our simulations confirm that otherwise bias between test forms in ability estimation arises.

Public Knowledge and Perceptions of Large-Scale Assessments: A Case Study

Yan Yan, Queen's University

This study examined public knowledge and perceptions about LSAs in a small Canadian province. A total of 515 questionnaires were completed. The overall findings revealed that public's perceptions towards LSA was in the middle of the scale and there were no statistically significant differences based on parental status, educational attainment, or cultural affiliations.

Differential Item Functioning in a Student Experiences Survey Across Contrasting University Pairs

Daniela Cardoza, The University of Iowa; Robert D. Ankenmann, University of Iowa; Thapelo Ncube, The University of Iowa

Measurement equivalence/invariance (ME/I) has been studied using confirmatory factor analysis (CFA) and differential item functioning (DIF). The purpose of this study is to compare ME/I DIF results to those of CFA in the 2016 Student Experiences in the Research University Survey across contrasting pairs of schools.

Item Response Theory Linking Methods for the Bifactor Model With Mixed-Format Tests

Sohee Kim, Oklahoma State University; Ki Matlock Cole, Oklahoma State University

This study compares IRT linking methods for the bi-factor model with mixed format tests. For the study, six different IRT linking methods are considered: four linking with linking coefficient (extension of mean/mean, mean/sigma, Haebara, and Stocking-Lord) and two linking without linking coeffi t methods (concurrent calibration and fixed item parameter calibration).

First-Order Learning Models With the Generalized DINA: Estimation With the Expectation Maximization Algorithm and Applications

Hulya Duygu Yigit, University of Illinois at Urbana-Champaign; Jeff Douglas, University of Illinois at Urbana-Champaign

The EM algorithm is presented for the estimation of student learning trajectories with the GDINA and some of its sub-models for the measurement component, and a first-order Markov model for learning transitions. A simulation study is conducted to evaluate estimation accuracy and an application using spatial reasoning data is given.

A Validity Argument Sensitivity Analysis of Social-Emotional Measures With Few Items

Carlos Chavez, University of Minnesota - Twin Cities; Michael C. Rodriguez, University of Minnesota; Julio Caesar Cabrera, University of Minnesota

Social and emotional learning is growing in interest among education researchers and practitioners. However, this interest is complicated by challenges in measurement, such as having few items. This paper investigates the challenge of using few items and the extent that model fit may be dependent on a single item.

Using a Treelike Item Structure to Disentangle Response Styles and Trait Information

Nikole Gregg, James Madison University; Brian Leventhal, James Madison University; Allison Ames Boykin, University of Arkansas

We disentangle an individual's response style from their substantive trait using a tree-like item structure. The two-stage decision-making process is represented with an IRTree model, where node-level item responses help determine whether the selection of 'Neutral' informs the substantive trait and response style trait estimates.

Measuring Learning Effectiveness: An Item-Level Dynamic Learning Model

Yanyan Tan, University of Georgia; Shiyu Wang, University of Georgia

This research generalized the Higher-Order Hidden Markov Cognitive Diagnostic Models [10] by allowing the benefit of practicing learning materials vary based on learning effectiveness parameters. A Bayesian estimation approach was evaluated through a simulation study. The proposed model is also applied to a spatial rotation data set.

Monday, April 8, 2019
10:25 – 11:55am, Manitoba, Coordinated Session

Keystroke Logs of Writing Processes in Large-Scale Assessments: Analyses and Applications

Chair: Mo Zhang, Educational Testing Service
Discussant: Sara Cushing, Georgia State university

There is a growing literature on the analytics and use of keystroke logs in digital writing assessments. The results and feedback based upon writing processes can not only enhance traditional score reports, which may contain only a single score, but also provide the test users with much targeted and rich information about the test takers. In this symposium, we will present the latest research related to the collection of keystroke logs and the reporting of writing-process features in large-scale assessments. These include the National Assessment of Educational Progress (NAEP), HiSET®, and a new K-12 standardized assessment at Educational Testing Service (ETS). An application of keystroke logging designed for test security purposes will also be presented.

Does Keyboarding Fluency Limit Writing Performance in Digital Writing Assessment?
Tao Gong, ETS; Gary Feng, Educational Testing Service; Mo Zhang, Educational Testing Service; Chris Agard, Educational Testing Service; Jie Gao, Educational Testing Service; Hilary Persky, Educational Testing Service; Patricia Donahue, ETS

Toward Understanding the Progression of Constructed-Response Fluency Across K–12 Age Groups
Daniel Adams, The University of Wisconsin - Madison; Jiangang Hao, ETS; Paul Deane, Educational Testing Service; E. Caroline Wylie, ETS; Elizabeth A. Stone, ETS; Gary Feng, Educational Testing Service

Measuring Writing Translation Using Keystroke Logs
Mo Zhang, Educational Testing Service; Jiangang Hao, ETS; Paul Deane, Educational Testing Service; Chen Li, ETS

Identifying Repeated Test-Takers Using Keystroke Information From Essay Writing
Paul Deane, Educational Testing Service; Jiangang Hao, ETS; Mo Zhang, Educational Testing Service; Ikkyu Choi, Educational Testing Service

Monday, April 8, 2019
10:25 – 11:55am, Quebec, Paper Session

Emerging Research on Longitudinal Diagnostic Classification Models

Discussant: Matthias Von Davier, National Board of Medical Examiners

Longitudinal Diagnostic Classification Modeling With Attribute Hierarchies

Wei Tian, Beijing Normal University; Jiahui Zhang, Michigan State University; Qian Peng, Collaborative Innovation Center of Assessment toward Basic Education Quality at Beijing Normal University; Haiyan Zhao, Beijing education examinations authority

Longitudinal diagnostic classification modeling (DCM) can be used to describe cognitive developmental trajectories in learning progressions. Previous studies have formulated models without attribute hierarchies. We incorporated attribute hierarchies via model constraints on the transition DCM and found the proposed model performed well under various conditions. A real-data application was provided.

Effects of Local Dependence on Longitudinal Diagnostic Classification Models

Yon Soo Suh, UCLA; Matthew James Madison, Clemson University

Measuring growth often involves using the same items multiple times and consequently, item-level dependencies can arise. This study investigates the impact of local item dependence for longitudinal diagnostic classification models (DCMs). We also explore potential remedies, including a hierarchical DCM to explicitly model such dependencies. Practical implications are discussed.

The Effects of Item Parameter Drift in Longitudinal Diagnostic Classification Models

Matthew James Madison, Clemson University; Laine Bradshaw, University of Georgia - Athens

Using a longitudinal diagnostic classification model (DCM), this study investigates the effects of item parameter drift (IPD) on classification accuracy and reliability. Simulation study results suggest that longitudinal DCMs are quite robust, able to provide consistently accurate and reliable classifications, even in the presence of substantial IPD.

Approaches to Estimating Longitudinal Diagnostic Classification Models

Junok Kim, University of California - Los Angeles; Matthew James Madison, Clemson University; Seungwon Chung, University of California - Los Angeles; Laine Bradshaw, University of Georgia - Athens

The utilization of diagnostic classifi tion models has recently expanded to longitudinal settings. This study compares and contrasts three different approaches to estimating longitudinal DCMs used in published studies: separate calibration, separate attributes for each time point, and latent transition analysis. Based on simulation results, theoretical and practical implications are discussed.

Monday, April 8, 2019
10:25 – 11:55am, Salon A, Coordinated Session

Detecting and Managing Testing Irregularities

Chair: Carol Eckerly, Educational Testing Service
Discussant: Gregory J. Cizek, University of North Carolina - Chapel Hill

Research on statistical methods to detect test fraud has become increasingly prevalent in recent years as the nature of the problem has changed, both in potential magnitude and the ways in which examinees can receive artificially inflated scores. This session presents three new methods which focus on various facets of test security, including detection of groups of examinees who collude or have access to a common key, examinees who harvest items or respond using item preknowledge, and items whose properties change over time due to item compromise. In addition to these new methodologies, this session explores aggregating information from a security investigation and communicating the results to relevant stakeholders. The session includes commentary and discussion from an expert in statistical detection of test fraud.

Answer Similarity Analysis at the Group Level
Carol Eckerly, Educational Testing Service

A CUSUM Procedure to Monitor Item Performance Over Time
Yi-Hsuan Lee, Educational Testing Service; Charles Lewis, Educational Testing Service

A New Person Fit Statistic for the Lognormal Model for Response Times
Sandip Sinharay, Educational Testing Service; Matthew Scott Johnson, Teachers College, Columbia University

When Is Fast Too Fast? Setting Score and Time Thresholds for Credentialing Programs
Angelica Rankin, Alpine Testing Solutions; Diane Talley, Alpine Testing Solutions; Jill van den Heuvel, Alpine Testing Solutions

Integrating Multiple Sources of Evidence in Test Security Analyses: Using Bayesian Inference to Weight the Strength of Evidence and Make Robust Decisions
William Skorupski, Amira Learning; James A. Wollack, University of Wisconsin - Madison; Sonya K. Sedivy, University of Wisconsin - Madison

Monday, April 8, 2019
10:25 – 11:55am, Salon B, Paper Session

New Insights in Differential Item Functioning Analyses

Discussant: Joni M. Lakin, Auburn University

RMSD: Limitations to Differential Item Functioning Detection in Low-Performing Populations

Yuan-Ling Liaw, University of Oslo; Maria Bolsinova, ACT, Inc.; David Rutkowski, Indiana University; Leslie Rutkowski, Indiana University; Jesper Tijmstra, Tilburg University

PISA 2015 implemented root-mean-square deviation for DIF detection. When the country-specifi profi y distribution matches the item location, the RMSD performs well. However, the RMSD is generally not well suited for detecting DIF in low-performing countries. We discuss the findings and the implications for future international assessment designs.

Using Explanatory Item Response Models to Reexamine Fairness in Psychometrics

Daniel Katz, University of California - Santa Barbara; Ronli Diakow, New York City Department of Education

This paper proposes a framework and statistical method for incorporating test-fairness analysis as part of standard psychometric validation work. A test-system perspective to fairness is proposed, focusing on a large-scale admissions test. This paper shows how explanatory item response modeling provides a method for this analysis.

Evidence of Fairness in Multilevel Data: AComparative Study of Three Differential Item Functioning Frameworks

Elizabeth Adele Patton, University of North Carolina - Greensboro

Previous research has examined multilevel DIF frameworks as solutions for nested data structure, however, there is a lack of research comparing across frameworks. A simulation study was conducted analyzing multilevel adjustments to the Mantel-Haenszel and SIBTEST and three-level Rasch model. Power, Type I error, and effect size estimates were compared.

Comparing Mantel-Haenszel and Wald Differential Item Functioning Detection Methods Under Matrix Sampling of Items

Xiaying Zheng, American Institutes for Research; Ummugul Bezirhan, Teachers College, Columbia University; Xinyu Ni, Teachers College, Columbia University; Seyfullah Tingir, Florida State University; Young Yee Kim, American Institutes for Research

This research compares the performance of the pooled-booklet Mantel-Haenszel procedure and the improved Wald tests in detecting DIF under item matrix sampling via a simulation study. Various data conditions are generated mimicking NAEP operational design. The performance of the two methods are evaluated based on power and Type-1 error.

What the MH DIF Statistic Is Designed to Measure

Hongwen Guo, ETS; Neil J. Dorans, ETS

The Mantel-Haenszel Differential item functioning (MH-DIF) procedure is currently the most widely used approach for assessing DIF in statewide assessment. However, many simulation studies found that it produced biased estimation of the latent- ability-based DIF criteria. We investigate what MH-DIF intends to measure theoretically, what factors cause bias, and the possible remedy.

Monday, April 8, 2019
12:20 – 1:50pm, Alberta, Invited Speaker Session

2018 NCME Career Award Session
Discussant: Derek C Briggs, University of Colorado - Boulder
Chair: Michael J. Kolen, The University of Iowa

A History of Classical Test Theory

Brian E. Clauser,
National Board of Medical Examiners

Monday, April 8, 2019
12:20–1:50pm, Algonquin, Coordinated Session

Examinations of Practices Used in Human Constructed Response Rating

Chair: Edward W. Wolfe, Educational Testing Service
Discussant: Kevin Raczynski, University of Georgia

This coordinated paper session of four papers and a discussant focuses on applied research that seeks to improve the human scoring enterprise by examining common practices from a new perspective to determine whether more effective and/or efficient procedures can be justified. Paper 1 (Attali) focuses on the rater certification process and critically examines the psychometric characteristics of those certification tests. Paper 2 (Wolfe) focuses on the information that is communicated to raters when they are monitored by attempting to determine what makes an essay difficult to score. Paper 3 (Wendler, Glazer, & Cline) focuses on the process of rater calibration and seeks to determine the optimal frequency of that practice. Paper 4 (Walsh, Arslan, & Finn) focuses on rater calibration and monitoring by examining whether the Predictive Performance Equation can account for raters' performances in these learning contexts. Kevin Raczynski of the Georgia Center for Assessment will serve as the Discussant for the session. Jointly, these four papers examine a range of practices across the entire duration of the human rating process with an aim of better understanding potentially causal relationships between rating practices and outcomes.

Rater Certification Tests: A Psychometric Approach
Yigal Attali, Educational Testing Service

Text Features of Difficult to Score Essays
Edward W. Wolfe, Educational Testing Service

Calibration Frequency and Its Impact on Scoring Accuracy
Cathy Wendler, ETS; Nancy Glazer, ETS; Frederick A. Cline, ETS

Computational Cognitive Modeling of Human Calibration and Validity Response Scoring for GRE and TOEFL
Matthew Walsh, RAND; Burcu Arslan, Educational Testing Service; Bridgid Finn, Educational Testing Service

Monday, April 8, 2019
12:20 – 1:50pm, Ballroom, Coordinated Session

Formative Assessment in the Disciplines: Advances in Theory and Practice

Chair: Gregory J. Cizek, University of North Carolina - Chapel Hill

This coordinated session has four main foci: 1) to present an overview of an emerging, comprehensive reconceptualization of formative assessment as consisting of both general foundations and discipline-specific elements and practices; 2) to describe how traditional and contemporary measurement perspectives and principles underlie the reconceptualization and provide a grounding for effective formative assessment; 3) to provide examples of discipline-based formative assessment practice in two areas—science and the arts. These areas were chosen intentionally as illustrative areas because, whereas some work is available in areas such as English Language Arts and Mathematics, less is available regarding the current redesign of science assessments in line with next generation science standards and the assessment challenges associated with this evolution, and because assessment in the arts also poses one of the most challenging contexts for applying innovative formative assessment practices. Fourth, reactions to the presentations, a synthesis of formative assessment practices in the disciplines, and suggestions for further evolution in theory, research, and practice will conclude the session.

New Conceptualizations of Formative Assessment in the Disciplines
Gregory J. Cizek, University of North Carolina - Chapel Hill

Integrating Measurement Principles Into Formative Assessment
Randy E. Bennett, Educational Testing Service

Formative Assessment Best Practices in the Disciplines I: The Arts
Heidi L. Andrade, University at Albany - SUNY

Formative Assessment Best Practices in the Disciplines II: Science
Erin Marie Furtak, University of Colorado - Boulder

Summary, Critique, and Future Directions
Dylan R. Wiliam, UCL Institute of Education

Monday, April 8, 2019
12:20 – 1:50pm, British Columbia, Coordinated Session

Testing Strategies, Extended Time Accommodation, and Speededness, Using Process Data in NAEP

Chair: Fusun Sahin, American Institues for Research
Discussant: Ryan Shaun Baker, University of Pennsylvania

In recent years, more assessments including the National Assessment of Educational Progress (NAEP) are transitioning from paper-based assessments (PBAs) toward digitally-based assessments (DBAs). The transition to DBAs permits the collection of detailed timing and behavior data on students' test taking behaviors. Automatically collected data on examinees' interactions with items and delivery interface during the test provide a rich data source to examine the relationship between students' testing behavior and performance from various aspects. This symposium features three separate studies investigating the relationship between students' testing behavior and performance, using the 2017 NAEP mathematics grade 4 (N=152,500) and/or grade 8 (N=148,100). DBA assessments administered to a nationally representative sample, respectively. All three studies analyzed two released blocks from each grade. The first study examines the relationship between students' time management strategies and performance. The second study examines effects of extended time accommodation, applying a propensity score matching approach. The third study deals with identifying rapid guessing, an issue frequently encountered in low stakes test, such as NAEP. This study uses growth mixture models (GMM) to identify rapid-guessers. Presentations in this symposium will show studies using process data can contribute to discourses on test assembly, test construction, and test validity issue.

Exploring Examinees' Timing Behaviors and Performance in a Digitally Based Mathematics Assessment
Fusun Sahin, American Institues for Research

Effects of the Extended Time Accommodation on Performance in NAEP Mathematics
Young Yee Kim, American Institutes for Research; Ruhan Circi, American Institutes for Research

Identifying Rapid-Guesser Using Growth Mixture Models
Xiaying Zheng, American Institutes for Research

Monday, April 8, 2019
12:20 – 1:50pm, Imperial Room, Electronic Board Session

Electronic Board Session 4

The Impact of Restrictive Models and Q-Matrix Misspecification on Classification Accuracy

Yanan Feng, Indiana University - Bloomington; Montserrat Valdivia; Dubravka Svetina, Indiana University - Bloomington; Justin Paulsen, Indiana University

This study examines whether using specific CDMs with restrictive assumptions affect the classification accuracy of examinees, compared to the more generalized LCDM framework. Since accurate Q-matrix specification is of great importance for CDMs. This study also aims at investigating whether Q-matrix misspecifi tion has differential impact on specific and general CDMs.

Generating Multivariate Data: Investigating Solutions Using Path Tracing Concepts

William R. Dardick, The George Washington University; Jeff R. Harring, University of Maryland - College Park

Using path tracing logic, we isolate a system of equations to solve for R-squared of one equation and another for covariance. The procedure uses SAS PROC MODEL (Jacobi method) and the NLPHQN function within SAS/IML (quasi-Newton method). We demonstrate data generation for multivariate regression models that motivate the study.

Estimating Conditional Standard Errors of Measurement Using a Multidimensional Item Response Theory Framework

Hacer Karamese, The University of Iowa; Won-Chan Lee, University of Iowa

It is often recommended that conditional standard errors of measurement (CSEMs) for different score units be reported if the standard error of measurement changes by score level. In this paper, multidimensional item response theory is considered as a framework for estimating CSEMs.

Relative Efficiency Diagnostics in Computer Adaptive Test Pools

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

This study evaluates relative efficiency index for a computer adaptive test (CAT). Pool level and conditional level relative efficiencies are summarized and compared with measurement precision indices and estimated scores. Results provide information on whether the index can serve as a diagnostic tool in developing effi t and parallel pools.

Impact of Weighted Sum Scores on Item Response Theory True Score Equating

Hyeonjoo J. Oh, ETS; Hongwen Guo, ETS

In this study, we compared the impact of two scoring methods (i.e., weighted sum score [WSS] vs. number correct score) on 2PL IRT true score equating results, particularly when a-parameters are lower than average. We also compared the IRT true score equating with WSS and IRT observed score equating.

A Comparison of Machine and Deep Learning Approaches in Automated Essay Scoring

Jinnie Shin, University of Alberta; Mark Gierl

The prediction accuracy and model behaviours of automated essay scoring frameworks developed with machine learning and deep learning algorithms were compared and thoroughly investigated. The results indicated that the deep learning-based AES could outperform the machine learning-based model producing more accurate and comparable results to the human-raters.

Evaluating Subgroup Differences on a Measure of Social and Emotional Learning

Yi-Lung Kuo, Beijing Normal University - Hong Kong Baptist University United International College; Alex Casillas, ACT, Inc.; Sonya J Powers, RTI International

ACT Engage is a measure of social and emotional learning (SEL) factors developed to be predictive of important academic outcomes. This study compares Engage score profiles across gender, ethnic/racial, and socio-economic subgroups, and evaluates whether the predictive validity of Engage differs across subgroups after correcting for range restriction.

Investigation of Different Insufficient Effort Responding Detection Methods in Educational Assessments

Nooree Huh, ACT, Inc.; Yu Fang, ACT, Inc.; Chi-Yu Huang, ACT, Inc.

The efficiency of different insufficient effort responding detection methods in online educational testing will be examined using students' response times (RT) and response patterns (RP). The examinees' person fi statistics and ability estimates will be compared among different subgroups that will be created based on the RT and RP.

A Small Sample Strategy to Initiate Pool-Based Professional Exams

Xinrui Wang, Pearson VUE

This study proposes a small sample strategy to initiate a pool-based exam program. This strategy aims at solving the practical concerns on scoring waiting period and item exposure in professional exam programs.

Investigating Internal Structure of Social Emotional Learning Measures: A Bifactor Approach

Mireya Carmen-Martinez Smith, University of Minnesota - Twin Cities; Youngsoon Kang, University of Minnesota - Twin Cities; Kory Vue, University of Minnesota - Twin Cities; Miranda Alejandra; Michael C. Rodriguez, University of Minnesota

Using the bifactor model, the internal structure of social emotional learning measures was investigated. Holding the general factor constant, we examined domain specific factors and factor loadings. Results suggest the measures have six domain specific factors and we recommend using all scores to predict external variables.

Moral Disengagement, Aggression, and Social Support: A Multiple Moderated Mediation Model

Yanni Shen, Beijing Normal University; Tao Xin, Beijing Normal University

This study identified the mediating role of anger and hostility in the relationship of moral disengagement and physical/verbal aggression. Results also found that the conditional indirect effect of moral disengagement in predicting physical/verbal aggression via hostility was weaker at high levels of social support than at low levels.

Comparing Item Response Prediction Based on Machine Learning and Explanatory Item Response Theory

Jung Yeon Park, University of Leuven; Frederik Cornillie, KU Leuven; Wim Van den Noortgate, KU Leuven

This study compares machine learning algorithms and explanatory IRT models to predict learner performance in online learning environments. Both approaches account for various features of the learners and items. A cross validation study using two educational data sets are demonstrated to identify the strengths and weaknesses.

Optimal Learning Strategy With Reinforcement Learning

Xiao Li, University of Illinois at Urbana - Champaign; Hanchen Xu; Jinming Zhang, University of Illinois at Urbana-Champaign; Hua-Hua Chang, Purdue University

We address the problem of determining an optimal learning strategy that can improve learning effi y for students on an E-learning platform, without prior information on the student learning model. A model-free reinforcement learning method is applied to solve the problem and a simulation validated the effectiveness of the proposed methodology.

Examining Predictive Algorithms for Licensure Exam Scoring That Are Trained on Unrepresentative Data

Christopher Runyon, National Board of Medical Examiners; Van Fan, National Board of Medical Examiners

Machine learning can be used to develop predictive models of expert ratings for written communication tasks on a medical licensure exam. We examine the performance of four regression models for predicting expert ratings when the training data is relatively small and unrepresentative of the examinee population.

Psychometric Properties of Difficulties in Emotion Regulation Scale With Latino University Students

Rui Jiang, University of California, Davis; Alexander Reid, California State University - Bakersfield; Bakhtiari Farin, UT - Austin; Scott Plunkett, California State University - Northridge

Confi matory factor analysis was used to examine the factor structure of four versions of the Diffi in Emotion Regulation Scale (36-item, 31-item, 18-item, 16-item) with Latino university students. Also, reliability and convergent validity were assessed. Each version had a good factor structure and appeared to be valid and reliable.

Acquisition of ELPA 21 BLV Form Item Parameters and Optimal Design

Sijia Huang, UCLA; Li Cai, University of California, Los Angeles

This paper (a) applies a mixed effect model to ELPA 21 2018 Judgment data to produce BLV form item parameters, (b) proposes an algorithm based on the D-optimality criterion to improve the design on judgment levels, (c) visualizes data and results to facilitate communications between test developers and users.

An Investigation of Polytomous Anchor Drift With Robust Z

Jungnam Kim, NWEA; Christie L. Plackner, DRC; Mayuko Simon, DRC; Dong-In Kim, DRC

This study examines polytomous anchor items and their multiple diffi y item parameters while using two Robust Z approaches to evaluate item parameter drift. The chosen approach can influence the estimated equating transformation values. Additionally, applying the Robust Z in an iterative or non-iterative manner is studied.

Transitioning From MTA to ATA: An Evaluation of Outcomes From Equating Methods

Kimberly Hudson, National Board of Osteopathic Medical Examiners

The purpose of this study is to evaluate outcomes from various Item Response Theory scale linking procedures and equating methods when Automated Test Assembly (ATA) was implemented. Upon estimating error, bias, and decision consistency indices, the mean/mean preequating method produced the most favorable results.

Validating Gf Measures Compiled via Automated Test Assembly

Jonathan P. Weeks, Educational Testing Service; Patrick Charles Kyllonen, Educational Testing Service

We compared the psychometric properties of four Gf tests, compiled via automated test assembly to results based on a validation sample of test takers who took all the measures. The results suggest that the test characteristics were adequately maintained when using the assembled forms.

Multidimensional Bifactor Modeling for Online Item Formats as Construct-Irrelevant Variance

Daeryong Seo, Pearson Assessment & Information; Se-Kang Kim, Fordham University

Several multidimensional models were hypothesized to measure item formats as construct-irrelevant variance (CIV) in different content domains. The multidimensional bifactor model including both item format and general ability factors was the best representation of the data. Both gender and ethnicity accounted for significant amount of CIV caused by item formats.

Motivational-Developmental Assessment for University Students: Generalizability Theory and Correlational Evidence

Brian F. French, Washington State University; David Alpizar, Washington State University - Pullman; Avi Kaplan, Temple University

Student motivational and developmental (MD) skills are critical to success and persistence in higher education. Innovative assessments are needed for understanding trajectories through postsecondary education. We examine a writing prompt-based MD measure for students through generalizability theory and associations with other variables. Results support score stability and relationships with outcomes.

National Council on Measurement in Education Invited Electronic Board 4

Andre A. Rupp, Educational Testing Service (ETS)

Monday, April 8, 2019
12:20 – 1:50pm, Manitoba, Coordinated Session

Developing Technology-Enhanced Items for Measuring Clinical Judgment in Nursing

Chair: Joe Betts, Pearson VUE
Discussant: April L. Zenisky, University of Massachusetts - Amherst

This coordinated session will highlight recent research on item development for measuring clinical judgment in entry-level nursing from initial concepts to formalization of item writing and review panels to field test results of item statistical functioning. Methods and processes presented are potentially useful for any program attempting to measure higher-order concepts. The first paper will discuss clinical judgment (CJ) in nursing and the overall assessment model conceived to drive development. The next paper will discuss the item development process from the initial panels to the final production of operationally ready items. Several item types will be introduced along with the suggested scoring approaches. The third paper will discuss the analysis and results of field test data gathered on the newly developed items on more than 30,000 examinees. The numerous score methods will be described along with an evaluation of the underlying dimensionality of the CJ construct. The final paper will provide research results related to scaling of the items using polytomous item response models for single items and item sets, e.g. testlets along with potentially useful methods of providing feedback on item performance to SMEs.

Developing a Task Model for Clinical Judgment in Nursing
Doyoung Kim, National Council of State Boards of Nursing; Joe Betts, Pearson VUE; William Joseph Muntean, Pearson

Developing Clinical Judgment Items From Task Model
Joe Betts, Pearson VUE; William Joseph Muntean, Pearson; Doyoung Kim, National Council of State Boards of Nursing

Evaluating Clinical Judgment Items: Field Test Results
Natalie Jorion, Pearson VUE; Joe Betts, Pearson VUE; Doyoung Kim, National Council of State Boards of Nursing; William Joseph Muntean, Pearson

Scaling Clinical Judgment Items Using Polytomous and Super-Polytomous Models
William Joseph Muntean, Pearson; Joe Betts, Pearson VUE; Doyoung Kim, National Council of State Boards of Nursing; Natalie Jorion, Pearson VUE

Monday, April 8, 2019
12:20 – 1:50pm, Quebec, Paper Session

Blossoming Research in IRTree Models

Discussant: Terry A. Ackerman, University of Iowa

Measuring Learning Outcome Using Responses and Response Times: Mastery and Fluency

Shiyu Wang, University of Georgia; Susu Zhang, University of Columbia

A general modeling framework of response accuracy and response times is proposed to track skill acquisition and provide additional diagnostic information on the fluency of applying the mastered skills in a learning environment. The proposed model is demonstrated through simulation studies and real data application.

Parameter Recovery of Two Item Response Tree Models: A Monte Carlo Simulation

Aaron Myers, University of Arkansas; Allison Ames Boykin, University of Arkansas

Item response tree (IRTree) models are often used to model response processes for personality scales, which can consist of few items, small sample sizes, and result in skewed data. This simulation evaluates the influence of these variables on parameter recovery for two commonly used IRTrees and provides recommendations for practitioners.

Effects of Category Labeling on Response Choice: An IRTree Analysis

Deborah L. Bandalos, James Madison University; Allison Ames Boykin, University of Arkansas; Elisabeth Marie Spratto, James Madison University

The issue of whether labels should be supplied for all numeric response options or only the endpoints of Likert-type items has received surprisingly little attention in the literature. We take an IRTree approach to this issue and show that a partially labeled scale resulted in more extreme responses.

Explaining Variability in Extreme Response Style Traits: A Covariate-Adjusted IRTree

Allison Ames Boykin, University of Arkansas; Aaron Myers, University of Arkansas

Extreme and midpoint response styles can confound the interpretation of scores and different respondent characteristics have been associated with response style. This study incorporates person-level covariates in an item response tree model for response style to explain the variability in response style. Bayesian estimation is used.

Building a Short Tree-Based Adaptive Screening Test for Juvenile Delinquency Risk

Yi Zheng, Arizona State University; Hyunjung Cheon, Arizona State University; Charles Katz, Arizona State University

In areas of screening for risks, the need for a short instrument is ubiquitous. Regression and computerized adaptive testing have been applied to shorten a long instrument. In this study, we develop a tree-based adaptive test for screening juveniles-at-risk following Gibbons (2013) based on a long questionnaire and real data.

Monday, April 8, 2019
12:20 – 1:50pm, Salon A, Paper Session

Innovative Applications of Machine Learning Techniques

Discussant: Susan Marie Lottridge, American Institutes for Research

Probing the Relationship Between Reading Profiles and Lexical Speech Features Using Machine Learning

Jeanne Sinclair, University of Toronto; Eunice Eunhee Jang, University of Toronto; Megan Vincett, OISE/University of Toronto; Hyunah Kim, University of Toronto; Samantha Dawn McCormick, University of Toronto; Christopher Douglas Barron, OISE/University of Toronto

Machine-learning and natural language processing (NLP) bring opportunities and challenges to language assessment. This study demonstrates and discusses the application of naïve Bayes and random forest algorithms using NLP-extracted lexical speech features to the prediction of reading comprehension profi that were generated through cognitive diagnostic modeling.

Evaluating Statistical and Machine Learning Methods to Improve Early Warning Systems

David Alexandro, Connecticut State Department of Education; Charles Martie, Connecticut State Department of Education; Christopher H Rhoads, University of Connecticut; Eric Loken, University of Connecticut; Suzanne M. Wilson, University of Connecticut; Hariharan Swaminathan, University of Connecticut

The purpose of this study is to evaluate statistical and machine learning methods to predict high school student performance and improve early warning systems (EWSs). The authors developed and compared the predictive accuracy of random forests, classifi tion and regression tree (CART, or decision tree), and regularized logistic regression models.

Forecasting Students' Future Academic Performance Using Big Data Analytics

Zhen Li, eMetric; Steven Tang, eMetric

Nowadays, more and more stakeholders make data-driven decisions. In education, data mining methods have become increasingly popular (Romero & Ventura, 2010; Zimmermann & et.al., 2015). This article compares two big data analytics methods, gradient boosted regression trees and Bayesian networks, for predicting students' future performance in state summative tests.

Monday, April 8, 2019
12:20 – 1:50pm, Salon B, Paper Session

New Directions in Scoring and Classification Accuracy

Discussant: Cristina Anguiano-Carrasco, ACT, Inc.

Classification Consistency and Accuracy for Tests With Conditional Dependence

Benjamin Andrews, ACT

A generalized multinomial error model is used to calculate classification consistency and accuracy for tests that violate conditional independence assumptions such as tests where a single response is evaluated with respect to several domains. Real data examples are presented along with simulations to evaluate the accuracy of different estimation procedures.

A Comparison of Four Scoring Methods for a Continuous Assessment

Xiaodan Tang, University of Illinois at Chicago; Andrew D Dallas, National Commission on Certifi ation of Physician Assistants; Fen Fan, National Commission on Certification of Physician Assistants; Joshua T. Goodman, National Commission on Certification of Physician Assistants

This study investigated a scoring problem posed by a continuous large-scale certification examination in which remediation items are rendered based on examinees' previous performance and survey questions about examinees' confidence, use of reference and content relevance. Four scoring methods were examined in terms of the estimation accuracy and classification consistency/accuracy.

An Examination of Classification Accuracy in a Continuous Testing Assessment Framework

Whitney Smiley, American Board of Internal Medicine

While testing organizations attempt to design tests to be as error-free as possible, there is undoubtedly error inherent within scores. While previous literature offers recommendations regarding controlling error, these concepts need to be reconsidered for different assessment frameworks. The paper outlines the impact of error in a continuous testing framework.

Integrating Expert Review and Diagnostic Classification Models for Online Assessments

Yuning Xu, SRI International; Mingyu Feng, WestEd; Daisy Wise Rutstein, SRI International; Wei Cui, Squirrel AI Learning

In this paper, we employed a method of combing expert review and diagnostic classification models (DCMs) to analyze the assessment in an online adaptive learning system. The intent was to investigate what happens to student classification if the alignment between the items and the latent skills was misspecified.

Standard Error of Variance Components, Measurement Errors, and Generalizability Coefficients in Single-Facet Generalizability Theory

Rashid S. Almehrizi, Sultan Qaboos University

Estimates of various variance components, measurement error variances, generalizability coefficients, like all statistics, are subject to sampling variability, particularly in small samples. Such variability is quantified traditionally through estimated standard errors and/or confidence intervals. The paper derives standard errors for all variance components, measurement error variances (relative and absolute), and generalizability coefficients for single-facet crossed design using delta method. A Monte Carlo simulation are performed for both normal data and dichotomously scored items with different test conditions. Results showed that the sampling variances for all estimators are converging to the true scores for both types of data.

Monday, April 8, 2019
2:15 – 3:45pm, Algonquin, Coordinated Session

Exploration of Issues With Applying Multistage Testing in NAEP

Chair: Xiaying Zheng, American Institutes for Research
Discussant: Mark D. Reckase, Michigan State University

This symposium presents research findings from three studies that address potential issues with implementing a multistage testing (MST) approach to the National Assessment of Educational Progress (NAEP), using the 2015 Grade 8 NAEP Mathematics pilot digitally based assessment (DBA) and the operational paper-based assessment (PBA). The first study examines if MST can increase the measurement precision and engagement. For engagement, this study compares engagement scores between linear and adaptive test versions with two randomly equivalent samples. Missing rates of cognitive items are also compared. For measurement precision, block level information is compared between the linear test and the adaptive test. The second study focuses on the issue of estimating ability in the routing stage, especially in mathematics with five subscales, by comparing the performance of three scoring approaches: UIRT, MIRT, and weighted average composite. The third study explores if time could be saved in the routing stage if all items in that stage were multiple-choice (MC) items. This study uses three sets of MC items only conditions: MC item only 12-item set, MC item only 15-item set, and mixed-format 18-item set. Presentations in this symposium will contribute to discourses on MST operational design in large-scale assessments, especially for NAEP.

Evaluation of MST Compared to Linear Assessment
Hyun Joo Jung, University of Massachusetts - Amherst; Young Yee Kim, American Institutes for Research; Soo Youn Lee, American Institutes for Research; Juliet Holmes, American Institutes for Research

Comparison of Weighted Average Composite Scoring Approach With Unidimensional Item Response Theory and Multidimensional Item Response Theory Approaches
Mingqin Zhang, The University of Iowa; Young Yee Kim, American Institutes for Research; Xiaying Zheng, American Institutes for Research

Can a Multiple-Choice-Items-Only Routing Block Reduce Routing Stage Time in MST?
Soo Lee, American Institutes for Research; Youngjun Lee, Michigan State University

Monday, April 8, 2019
2:15 – 3:45pm, Ballroom, Coordinated Session

Formative Multimodal Assessment of Collaborative Problem-Solving Skills

Chair: Yigal Rosen, ACTNext and Harvard University
Discussant: Saad Khan, ACTNext

Collaboration among peers is a common practice in workplace and learning environments. Typically, peers establish shared understandings of the problem space and their expertise, divide workload and responsibilities, take actions to advance objectives, monitor progress, and provide feedback. Numerous research reports indicate that collaborative problem solving is increasingly important in today's complex interconnected world, therefore of increasing interest for teaching and assessing with students (OECD, 2017). However, assessment design, scoring, data analytics and reporting on CPS skills, specifi in the context of formative multimodal assessment, is challenging. In this symposium a spectrum of approaches for CPS assessments and data analytics will be introduced, and four papers will be presented and discussed.

Measuring Learning Gains in Centralized Versus Distributed Collaborative Problem-Solving Groups
Yigal Rosen, ACTNext and Harvard University; Iris Wolf, World ORT Kadima Mada

Developing Scalable Conversational Assessments for Collaborative Problem Solving
Yigal Rosen, ACTNext and Harvard University; Kristin Stoeffler, ACT, Inc.

Exploring CPS Skill Evidence and Behavior Models Using Machine Learning Analytics
Pravin V Chopade, ACT, Inc.; David Edwards, ACTNext; Spencer Swartz, ACTNext; Saad Khan, ACTNext

Using Technology-Rich Environments to Explore Gender Differences During Collaboration
Iris Bourgault-Bouthillier, Université de Montréal; Kristin Stoeffler, ACT, Inc.; Yigal Rosen, ACTNext and Harvard University; Alina A. Von Davier, ACT, Inc.

Monday, April 8, 2019
2:15 – 3:45pm, British Columbia, Paper Session

Advances in Multidimensional Item Response Theory

Discussant: Michael Toland, University of Kentucky

Estimating Students' Topic-Level Abilities Using Extended Higher Order Item Response Theory Models

Weimeng Wang, University of Maryland, College Park; Jie Sun, American Institutes for Research; Hong Jiao, University of Maryland-College Park

In the large-scale assessment, subscores are usually reported at the domain level. However, the general information reported at the domain level is not suffi t for providing more fi -grained information for learning and instruction. This study explores two extended higher-order IRT models to estimate subdomain ability at the topic level.

A Multidimensional Hierarchical Framework for Modeling Speed and Ability

Peida Zhan, Zhejiang Normal University; Hong Jiao, University of Maryland-College Park; Wen-Chung Wang, University of Hong Kong, Hong Kong; Kaiwen Man, University of Maryland - College Park

In educational multidimensional tests, latent speed may also be multidimensional. This study fi proposed a multidimensional log-normal response time (RT) model to consider the multidimensionality of latent speed. Further, to simultaneously take into account the response accuracy and RTs, a multidimensional hierarchical modeling framework was proposed.

Multidimensional Test Assembly of Parallel Test Forms Using Mixed-Integer Linear Programming

Dries Debeer, University of Zurich

The statistical target commonly used for the assembly of parallel test forms in unidimensional IRT is not directly transferable to multidimensional IRT. Therefore several new statistical targets are proposed, some of which are Kulback-Leibler information indexed (KLI). The different approaches are compared and evaluated in uni- and multidimensional cases.

Estimation of Multidimensional Item Response Theory Models Using Higher Order Asymptotic Expansions

Björn Andersson, University of Oslo

We present a method for estimation of multidimensional IRT models using a second-order Laplace approximation. The method is highly computationally efficient for IRT models with a simple structure. In a simulation, the method has lower bias and mean squared error compared to alternatives while being substantially faster.

Monday, April 8, 2019
2:15 – 3:45pm, Imperial Room, Electronic Board Session

Electronic Board Session 5

Classification Predictors for Reading Fluency in Oral Reading Fluency Assessment

Akihito Kamata; Yusuf Kara, Southern Methodist University; Chalie Patarapichayatham, Southern Methodist University; Thu Le, Southern Methodist University

This study investigates potential predictors of reading fluency in student oral reading fluency assessment passages. We derive a number of variables from word-level reading time, silence time, and accuracy. Then, we fit machine learning algorithms to explore how the derived variables classify readers into groups of different levels of fluency.

An Examination of Item Calibration Methods in Multistage Testing

Liuhan (Sophie) Cai, Measured Progress; Louis A. Roussos, Measured Progress

This study uses a large-scale MST-administered mathematics assessment dataset and compares the pre-equating model to post-equating models based on different item calibration methods in terms of item parameter estimates, person parameter estimates, and classification in the realm of MST.

Considerations in S – χ2: Rest or Summed Score, Priors, and Violations of Normality

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

The S-χ2 item fit index is one of the few item fit indices that appears to maintain accurate Type I error rates. This study explored grouping examinees by the rest score or summed score, prior distributions for the item parameters, and the shape of the ability distribution.

Differential Item Functioning Analysis for Immigrant Status for the 2015 PISA Science Items

Gonca Usta, SMU; Akihito Kamata, Southern Methodist University

The purpose of this study is to use Differential Item Functioning (DIF) analysis to investigate differences in the performance of immigrant and non-immigrant students in 48 cognitive science items for the U.S. sample in 2015 PISA. According to the results, all items displayed level-A DIF, indicating that the items had negligible DIF effect.

Differential Item Functioning Item Detection With Different Q-Matrices Under CDM

Hueying Tzou, National University of Tainan; Pei-Ming Chiang, National University of Tainan; Yi-Fang Wu, ACT, Inc.

International large scale assessments are used to compare student performances across countries. However, same items might measure different skills because learning strategies or instructions given differ by country. Under CDM, this study aims to investigate whether DIF detection procedures can identify items that measure different attributes across countries.

Nominal Response Model to Address Missingness in Multistage Adaptive Testing

Dee Duygu Cetin-Berber, University of Florida; Okan Bulut, University of Alberta; Corinne Huggins-Manley, University of Florida

This study investigates the performance of multistage adaptive testing in presence of missing responses. Nominal response model (NRM) is utilized to treat missing responses as a separate response category. Initial results showed that NRM provided unbiased (e.g., bias < 0.05) theta estimates with 30% of missingness in 1-2-3 MST design.

Modeling of Multilevel Item Structures: A Comparison of Item Response Theory Calibration Strategies

Richard M. Luecht, University of North Carolina - Greensboro; Elizabeth Adele Patton, University of North Carolina - Greensboro; Alexandra Lay, University of North Carolina - Greensboro

This large-scale simulation study compares two analysis strategies for nested item structures: multilevel analysis and multi-stage single-level analysis with data restructuring. Multiple item-production conditions are investigated in both educational achievement and certification/licensure testing contexts. The results are germane for QC in operational testing settings where automatic item generation is implemented.

A Local Differential Item Functioning Method to Evaluate Constructed Response Scoring Shift With Simulated/ Real Data

Xuan (Adele) Tan, ETS

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

Moderated Regression Variants via Theoretically Meaningful Constraints to Assess Subpopulation Differences

Ernest C. Davenport, Jr., University of Minnesota; Kyungin Park, University of Minnesota - Twin Cities; Mark L. Davison, University of Minnesota

The proposed study uses of variants of moderated regression to develop models via constraints that are testable and theoretically meaningful. These models allow us to parse the relationship between criteria, predictors, and groups to address achievement gaps. This approach will be demonstrated on data from NCES.

Measuring Subpopulation Noninvariance in Test Equating: How Much Is Too Much

Ian Campbell, University of Notre Dame

The RESD statistic is a common ways to measure the amount of subpopulation non-invariance in test equating, but its behavior for finite sample sizes is relatively unknown. Through equivalence testing simulations and data analyses, we demonstrate the utility of different threshold levels for evaluating how much non-invariance is too much.

The Influence of Rating Scale on Coefficient Alpha and Alternative Reliability Measures

Guher Gorgun, University at Albany - SUNY; Kimberly F. Colvin, University at Albany - SUNY

In this study, coefficient alpha and several alternative reliability measures were compared. The Rosenberg self-esteem scale (Rosenberg, 1965) was administered with different rating scales. Tau-equivalence, test of homogeneity, standardized factor loadings and reliability coefficients were analyzed across different versions of rating scales used. Practical implications are discussed.

Exploring Item Difficulty in Assessments of Computer Programming With Cognitive Interviews

Matt Davidson, University of Washington - Seattle; Dongsheng Dong, University of Washington - Seattle; Min Li, University of Washington; Benjamin Xie, University of Washington; Andrew Ko, University of Washington

Research on introductory computer science assessments has focused on construct validity. This study investigates item features that contribute to difficulty through cognitive interviews with university students. Transcripts were analyzed alongside a framework of item features. Findings have implications for validity arguments about response processes as well as item writers.

Admission Testing Through Curriculum-Sampling: Easy to Communicate and High Predictive Validity

Rob R. Meijer, rijksuniversiteit groningen; Anna Susanna Maria Niessen, University of Groningen

We investigated the validity of curriculum-sampling tests that mimic representative parts of academic programs to predict future academic achievement in higher education. The curriculum-sampling tests showed high predictive validity for first- and third-year academic achievement, incremental validity over high school GPA, and was related to perceived test competence.

Four Methods for Assessing Measurement Invariance in Many Groups: PISA Teacher Support

Carina McCormick, Buros Center For Testing

The study compares the results of four methods for assessing measurement invariance (MI) with many groups: traditional CFA MI, Bayesian approximate MI testing, alignment, and alignment-within-CFA (AwC). The study focuses on the six-item Teacher Support scale for mathematics teaching, using data from PISA 2012, with analyses conducted in Mplus.

An Idiographic Perspective on Dimensionality Using Person Response Functions

Victoria Tanaka, University of Georgia - Athens; George Engelhard, University of Georgia

This study examines within person dimensionality using person response functions. Current discussions of dimensionality are nomothetic and group-based, while this study explores an idiographic and individual-level perspective based on how each person responds to items. An idiographic perspective reveals potential person misfit that can be diagnosed using person response functions.

The Impact of Test-Taking Disengagement on Item Content Representation

Steven L. Wise, Northwest Evaluation Association

Disengaged test taking, as identified through rapid-guessing behavior, degrades score validity in several ways: it tends to negatively distort scores, it implies overestimated score precision, and it distorts the intended content representation of engaged responses. This paper discusses and empirically demonstrates each of these psychometric costs associated with rapid guessing.

Granular Individual-Level Growth Modeling via Bayesian Network Analysis

Jinah Choi, Edmentum, Inc.; Windy Torgerud, Edmentum, Inc.

This study presents an individual-level growth model fueled by digital learning data collected in classrooms across the country. Mastery quizzes, performance checks, and computerized adaptive testing information feed into a Bayesian network integrating these elements to inform domain specifi and overall subject knowledge representation of growth over time.

How Are Above-Grade Items Useful in Vertical Scales?

Duy Pham, University of Massachusetts, Amherst; Brian Francis Patterson, Curriculum Associates, LLC; Kevin James Cappaert, Curriculum Associates; Daniel F. Mix, Curriculum Associates

In this study, we explored the usefulness of using above-grade items to assess high performers in a vertically scaled computer adaptive assessment of mathematics proficiency. Our findings signified that using these items helped us improve the accuracy of categorizing examinees into the top 5%. Implications and future directions are discussed.

Psychometric Effects of Technology-Enhanced Item Scoring and Proportions in Computer Adaptive Testing
Chunmei Zheng, Pearson; David Shin, Pearson

Growing interest has been expanded in developing and deploying technology enhanced items. No research studies have explored psychometric properties of a test with different scoring and proportion for technology-enhanced items in computer adaptive testing. This study will investigate the psychometric effects of different scoring and proportion for technology-enhanced items.

Bayesian Model Checking in Cognitive Diagnostic Models
Nan Wang, Florida State University

Checking that models adequately present data is an essential component of applied statistical inference. The purpose of this dissertation is to use prior predictive posterior simulation method and posterior predictive method to investigate the person fi of DINA model with chosen discrepancy measures of log-likelihood statistic and unweighted between-set index.

National Council on Measurement in Education Invited Electronic Board 5
Andre A. Rupp, Educational Testing Service (ETS)

Monday, April 8, 2019
2:15 – 3:45pm, Manitoba, Paper Session

Advances in Item Development, Pretesting, and Selection

Discussant: Joshua T. Goodman, National Commission on Certification of Physician Assistants

Predicting Item Survival Using Natural Language Processing

Victoria Yaneva, National Board of Medical Examiners; Peter Baldwin, National Board of Medical Examiners; Janet Mee, National Board of Medical Examiners; Le An Ha, University of Wolverhampton

Being able to predict the survival rate of items that are being pretested can significantly reduce cost and effort. We provide evidence for differences in the way surviving and non-surviving items are written by combining 113 automatically extracted linguistic features in a machine-learning model aiming to predict item survival.

Automatic Enemy Item Detection Using Natural Language Processing

Fang Peng, University of Illinois at Chicago; Kimberly A. Swygert, National Board of Medical Examiners; Ian Micir, NBME

Latent Semantic Analysis (LSA) offers computational methods for extracting and representing the meaning of words as underlying dimensions of a large text corpus (Landauer and Dumais, 1997). This paper presents an automatic approach of using LSA to measure item similarity with the goal of identifying enemy relationship in item pools.

A Revised Index for Detecting Nonfunctional Distractors in Multiple-Choice Questions

Mark R. Raymond, National Board of Medical Examiners; Craig Stevens, National Board of Medical Examiners

MCQ distractors are traditionally defi as "nonfunctional" if selected by fewer than 5% of examinees. This definition is problematic for easy items because there are few unknowledgeable examinees available to choose the distractors (e.g., mastery tests). This paper introduces an index of nonfunctional that is sensitive to item difficulty.

Item Readability Attributes as Drivers of Operational Success

Matthew Schultz, Association of International Certified Public Accountants; Joshua Stopek, Association of International Certified Public Accountants; Seung-Hee (Sam) Chung

Content development costs are typically significant. Luecht (2005) notes that the average cost-per-item (ACPI) can range to over $1500 per item. The ability to predict pretest items likelihood of successful calibration can have significant implications. Results suggest that readability metrics can provide useful feedback to content developers regarding operational success.

Analyses of Distractors in English Summarizing Test Items: Focusing on Cognitive Processes

Takahiro Terao, The National Center for University Entrance Examinations; Hidetoki Ishii, Nagoya University

This study, using an originally developed test, aimed to compare attractiveness of distractors by their proficiency, based on Japanese students' typical errors in writing a summary of an English paragraph. Distractors presented as summaries lacking essential information and without the author's viewpoints were more attractive in less proficient students.

Monday, April 8, 2019
2:15 – 3:45pm, Quebec, Paper Session

Issues and Challenges With Adaptive Testing

Discussant: Carl Setzer, AICPA

The Relationship Between Item Bank Structure and Population Distribution for Noncompensatory MAT

Chia-Ling Hsu, The Education University of Hong Kong

As the usefulness of MAT and the complications associated with non-compensatory data. We investigated the relationship between the distributions of the administered items and examinees, furthermore, to shed light on item bank construction. Simulations showed that the mean of bank distribution would shift to the left compared with population distribution.

Impact of Collateral Information on Ability Estimation in an Adaptive Test Battery

Qing Xie, The University of Iowa; Deborah Harris, University of Iowa; Terry A. Ackerman, University of Iowa; Catherine Welch, University of Iowa

The purpose of this study is to compare different ways of incorporating collateral information under the unidimensional and multidimensional CAT frameworks and to investigate the impact of subtest intercorrelations and sequences of subtest administration on ability estimation in a variable-length adaptive battery with content constraints and item exposure control.

Measurement Precision and Efficiency of Multistage Tests and Computerized Adaptive Tests

King-Yiu Suen, University of Minnesota; David J. Weiss, University of Minnesota - Twin Cities

This study compared the measurement precision and efficiency of computerized adaptive testing and multistage testing (MST). The influences of three MST design factors, namely, test structure, item allocation, and assembly priority, were also investigated. The impact of routing errors on MST performance was also evaluated.

Impact of Random Guessing in Linear and Adaptive Tests

Briana Hennessy, University of Connecticut; Eric Loken, University of Connecticut; Jennifer Richardson, University of Connecticut

Low test taker motivation adversely affects ability estimation. We compare the impact of random guessing behavior on a linear test with ordered items and a computer adaptive test. The error and bias induced by giving up part way through the test is very different across test type.

Monday, April 8, 2019
2:15 – 3:45pm, Salon A, Paper Session

Issues, Opportunities, and Challenges With Cognitive Diagnostic Modeling

Discussant: Laine Bradshaw, University of Georgia - Athens

The Impact of Conditional Dependency and Its Detection on Cognitive Diagnosis Modeling

Kevin Carl Pena Santos, University of the Philippines; Jimmy de la Torre, The University of Hong Kong

Estimating cognitive diagnosis models (CDMs) typically assumes conditional independence. This study investigates how violating this assumption affects the validity of inferences obtained from CDMs, and how it can detected. Results indicate that conditional dependency yields biased estimates and understated standard errors, resulting in higher classification accuracy and item discrimination indices.

A General Diagnostic Classification Model for Rating Scales

Ren Liu, University of California - Merced; Zhehan Jiang, The University of Alabama - Tuscaloosa

This study proposes and evaluates a general diagnostic classification model for rating scales. Findings suggest that the proposed model shows promise for accommodating much smaller sample sizes by reducing a large number of parameters for estimation and obtaining similar item category response probabilities and individual scores with a saturated model.

Cognitive Diagnostic Modeling With Hierarchical Attributes: Full Q-Matrix Versus Reduced Q-Matrix

Jiahui Zhang, Michigan State University; Qian Xu, Michigan State University

Short tests based on cognitive diagnostic modeling with hierarchical attributes provide useful psychometric tools for formative classroom assessments. We discuss parameterizations and Q-matrices for three models under linear, divergent and convergent hierarchies. Differences between reduced and full Q-matrices and their effects on attribute profile classification were explored in simulation studies.

Linking Diagnostic Systems for Mathematics Deficits to a Dynamic Learning Map

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

This study examines linking a diagnostic system of mathematical defi (Embretson, 2015) to the Enhanced Learning Map (ELM; Kingston & Broaddus, 2015) to permit selecting instruction for students' mathematical deficits. ELM is a broad network of standards-based knowledge and skills that accommodates national and international standards.

Examining the Performance of a Neural Network Cognitive Diagnostic Model in Hierarchy-Structured Attributes

Chi Chang

Cognitive diagnostic models require large sample size, which makes their application in the classroom setting impossible. This study applied a neural network approach to a cognitive diagnostic model and examined its performance under 48 scenarios. The results showed that the proposed approach can better handle small sample size and hierarchically-structured attributes.

Monday, April 8, 2019
2:15 – 3:45pm, Salon B, Paper Session

Important Considerations in Setting Cut Scores

Discussant: Jonathan Beard, The College Board

Setting Instructionally Informative Cut Scores in a Formative System

Karen Barton, Edmentum; Jinah Choi, Edmentum, Inc.; Karla Egan, EdMetric; Anne H. Davidson, EdMetric LLC

Standard setting methods operationalize performance expectations of content through systematic processes. This study compares two methods for setting cut scores on a system of formative assessments aligned to learning progressions by articulating and comparing state-level performance level descriptors (PLDs) in terms of 1) instructional content and 2) assessment data.

Setting Multiple Standards on a Vertically Scaled Multistage-Adaptive Test

Jennifer Lee Lewis, University of Massachusetts - Amherst; Hwanggyu Lim, University of Massachusetts; Frank Padellaro, Measured Progress; Stephen G. Sireci, University of Massachusetts - Amherst

What is the best method for computing cut-scores for a vertically scaled, multistage-adaptive test (MST)? Currently, there is limited guidance regarding an appropriate method to establish cut-scores on MSTs. This paper illustrates and evaluates three methods of computing cut-scores and provides evidence to support the validity of the cut-scores.

Pass-Rate Stability for Repeated Standard-Setting Versus Equating With Small Samples

Patrick Meyer, Northwest Evaluation Association; Valerie Link, CFA Institutute

Pass rates are typically obtained by equating a new form to one used during standard setting. An alternative is to conduct standard setting on every new form and compute pass rates directly. We compared these methods using data from seven administrations of a certification exam given over a five-year period.

A Comparison of Two Methods of Setting Standards for an Adaptive Test.

Steven J. Fitzpatrick, Pearson; David Shin, Pearson; Mary Kino, Pearson

Two approaches to setting standards on a computer adaptive mathematics test are compared. The fi is a modified Angoff Yes/No method with a fixed set of items. In the second, panelists take the CAT test and responded in a manner reflecting borderline performance. Cut score and survey results are presented.

Monday, April 8, 2019
4:10 – 6:10pm, Alberta, Paper Session

Challenges, Issues, and Opportunities in Interrater Reliability

Discussant: Jason P. Kopp, American Board of Surgery

A New Test of Rater Drift in Trend Scoring

John R. Donoghue, Educational Testing Service; Carol Eckerly, Educational Testing Service

Trend scoring (rescoring Time A at Time B) constructed response items gives rise to two-way data that follow a product multinomial, rather than the usual multinomal distribution. A statistic, based on comparing conditional distributions, is introduced. A simulation examines distributional properties and compares performance with paired t-test and Stuart's Q.

Using Rater Cognition to Improve Generalizability of an Assessment of Scientific Argumentation

Katrina Borowiec, Boston College; Courtney Castle, The Woodrow Wilson Academy of Teaching and Learning

To improve the quality of scoring constructed response items, cognitive interviews provide information about raters' scoring judgments. Rubrics for scientific argumentation items were modified based on rater interviews. A g-theory analysis was conducted to measure change in interrater reliability, accuracy, and generalizability. All three measures increased with the rubric modifications.

Relationship Between Intraclass Correlation and Percent Rater Agreement

Jason Bryer, Excelsior College; Guher Gorgun, University at Albany - SUNY

Inter-rater reliability is critical for establishing reliability. Current literature suggests that intraclass correlation coefficients (ICC) be used over percent rater agreement (PRA), however, interpreting ICC is difficult. This article explores the relationship between ICC and PRA using simulations; results indicate they are highly correlated ($R^2 > 0.85$) for most designs.

Rater Consistency With a Teacher Observation Protocol

Evelyn Johnson, Boise State University; Yuzhu Zheng, Boise State University; Angela Rae Crawford, Boise State University; Laura Moylan, Boise State University

We investigated raters' application of the scoring criteria of an observation protocol using many-faceted Rasch measurement (MFRM) and qualitative "think aloud" analysis. Analyses showed that while raters scored in an internally consistent manner, they differed in their severity, interpretation of items, and identification of supporting evidence for their scores.

Using Standard-Setting Methods for Length and Original Text Thresholds in Essays

Ahmet Turhan, American Institutes for Research; Sue Lottridge, American Institutes of Research; Julie Benson, American Institutes of Research; Jon Cohen, American Institutes for Research

We outline the methods and results of two threshold-setting events held with writing educators in an eastern state using standard setting-like methods. The results showed that word count thresholds that varied across grades and proportion copied thresholds that were similar across grades.

Monday, April 8, 2019
4:10 – 6:10pm, Algonquin, Paper Session

Assessing Dimensionality: Emerging Research and Technical Considerations

Discussant: Jessalyn Smith, Data Recognition Corporation

Assessing Dimensionality Due to Item Types

Nicole Zelinsky, University of California - Merced

Exams should be modeled at the correct level of dimensionality to avoid biased parameters and miscategorization of examinees. This study assessed dimensionality of a high-stakes medical licensure exam due to item types. Models did not suggest individual dimensions for each item type but showed dimensionality due to local-item dependence.

Classifying Subdomain Performance in a Variable-Length Computer Adaptive Test

Chen Li, Kaplan Test Prep; Michael Chajewski, Kaplan Test Prep

When using subdomain ability estimates and their conditional standard errors of measurement (CSEM) to classify subdomain performance, inconsistency can occur between test and subdomain classifications for a variable-length computerized adaptive practice test. This study investigates two classification methods to improve the consistency between the overall test and the subdomain classifications.

Statistical Dimensionality of Standardized Test Forms by Examinee Course-Taking Behavior

Alexandra Lay, University of North Carolina - Greensboro; Terry A. Ackerman, University of Iowa

Statistical dimensionality is a function of test and examinee characteristics. Because standardized tests are often administered to a diverse examinee population, an investigation into characteristics that affect test dimensionality is warranted. This study aims to explore the extent to which examinee course taking behavior affects statistical dimensionality.

Exploring Factors Affecting Subscore Reporting in Multistage Adaptive Testing

Yanming Jiang, ETS

This study examines factors that influence the reliability of subscores and the accuracy of subscale ability estimates in multistage adaptive testing. The factors considered in the study include subtest length, correlations among subscores, and item pool characteristics such as the number of items and their statistical properties.

Dimensionality Assessment With Locally Dependent Item Responses: Scoring Rules Versus Raw Data

Alex Brodersen, University of Notre Dame; Qian Hong, NCSBN; Doyoung Kim, National Council of State Boards of Nursing

Dimensionality assessment of locally dependent (LD) responses (e.g. technology enhanced items /testlets) is explored. Applying scoring rules or analyzing the raw data results in over or under estimating the number of latent dimensions. New methods were developed and evaluated to accommodate LD in dimensionality detection without applying scoring algorithms.

Monday, April 8, 2019
4:10 – 6:10pm, Ballroom, Paper Session

Assessment as Feedback for Teachers and Students

Discussant: Kristen Huff, Curriculum Associates

Insights Into Editing and Reviewing in Writing Process Using Keystroke Logs

Mengxiao Zhu, ETS; Mo Zhang, Educational Testing Service; Paul Deane, Educational Testing Service

Using the keystroke logs from 761 middle school students in the US, this study started from the character-level inputs and reconstructed the sentence-level actions to capture the editing and reviewing behaviors in the writing process. Preliminary findings showed different writing behavior patterns for participants with different scores and demographic features.

Reporting Results From Diagnostic Classification Models for Teachers

Zachary Feldberg, University of Georgia; Laine Bradshaw, University of Georgia - Athens

To maximize the utility of diagnostic classification models (DCMs), reports must enable educators to accurately discern student competencies. We examine teacher interpretations of results from DCMs to investigate the degree to which they are appropriate for DCMs and provide recommendations for designing DCM reports to facilitate teacher understanding.

Teacher Assessment Literacy: Implications for Diagnostic Assessment Systems

Amy Clark, The University of Kansas; Brooke Nash, The University of Kansas; Meagan Karvonen, The University of Kansas

Assessment literacy centers on teachers' basic understandings of fundamental measurement concepts and their impact on instructional decision-making. The rise of diagnostic assessment systems has important implications for teachers' assessment literacy and their interpretation and use of assessment results. This study examines teachers' assessment literacy in a diagnostic assessment context.

Digital Platform Instructor-Facing Learning Analytics: Theoretical Basis of Design and Instructor Perceptions

Kimberly Rebecca Marsh Runyon, Macmillan Learning; Billie-Jo Grant, Macmillan Learning; Erin Scully, Macmillian Learning; Lisa Ferrara, Macmillan Learning; Kara McWilliams, Macmillan Learning

The current study presents an evidence-based approach to the design of instructor-facing learning analytics within a higher education digital learning solution. A series of mixed methods formative evaluation studies were conducted to examine instructor perceptions of preliminary dashboard designs. Implications for higher education practice and digital platform enhancements are discussed.

Exploring the Effect of a Scaffolding Design in Argument Critique Tasks

Yi Song, Educational Testing Service; Szu-Fu Chao, Educational Testing Service; Yigal Attali, Educational Testing Service

This project aimed to examine the effect of scaffolding on student performance on argument critique tasks. We administered the tasks to 472 students from three middle schools that have high proportions of students from low-income families. We present the tasks, scaffolding design, study results, and research implications.

Monday, April 8, 2019
4:10 – 6:10pm, British Columbia, Paper Session

Important Test Administration and Scoring Considerations

Discussant: Michael B. Bunch, Measurement Incorporated

The Effects of Test Familiarity on Person-Fit and Aberrant Behavior

Hotaka Maeda, University of Wisconsin-Milwaukee; Xiaolin Wang, NBOME

The person-fi to the Rasch model was evaluated for examinees taking multiple subject tests with a similar structure. The evaluation considered which test in the sequence (i.e., first, second) was taken. Compared to an examinee's first test, person-fit improved for later tests. Test score reliability may improve with test familiarity.

Digital Familiarity and Warm-Up Effects in NAEP 2017 Mathematics

Markus Broer, American Institutes for Research; Ruhan Circi, American Institutes for Research; Young Yee Kim, American Institutes for Research

In digital based assessments (DBAs), students with low digital familiarity may use overall more time, but particularly early in the assessment (warm-up effect), which may or may not impact their performance. This study investigates the relationship between digital familiarity, warm-up effects and performance in a NAEP DBA.

Response Time Analysis for Indicating Low Student Motivation on a Summative Assessment

Stefan Jansen, CITO; Hendrik Straat, Cito; Zijlstra Wobbe, Cito; Daniel van der Palm, Cito

This study aims to improve the communication about do's and don'ts in digital assessments. We use different perspectives on a combination of response time and accuracy, and prior information to demonstrate that response time affects the result. Results show an effect of response time on the response accuracy.

On Administering Salt Items in Computerized Adaptive Testing With Salt

Zhongmin Cui, ACT Inc; Chunyan Liu, National Board of Medical Examiners; Yong He, ACT, Inc.

Computerized Adaptive Testing with Salt (CATS) implements CAT with unrestricted item review and answer changes while being robust to cheating strategies. A new algorithm on administering salt items under CATS was proposed to improve test efficiency. The new algorithm was shown to perform better than the original CATS.

Monday, April 8, 2019
4:10 – 6:10pm, Manitoba, Paper Session

Assessment of Special Populations and Subgroups

Discussant: Hyeonjoo J. Oh, ETS

Validating Learning Pathways for Students With Significant Cognitive Disabilities

Yuxi Qiu, University of Florida

In this study, Bayesian networks are utilized to portray and further to validate the learning pathways for students with significant cognitive disabilities on the basis of the Dynamic Learning Maps Alternate Assessment. Results from model evaluation and the associated implications are discussed.

Read-Aloud Accommodations: Moving Toward a Well-Reasoned Approach to Investigation and Provision

Sara Witmer, Michigan State University

Research on the read-aloud accommodation frequently ignores the heterogeneity of students with disabilities. This study explores the measurement comparability of academic content area tests administered with and without a read-aloud accommodation specifically to students with word recognition and word study difficulties. Implications for policy and practice will be discussed.

Trends in Participation, Performance, and Accommodations Received by Special Education Students

Yi-Chen Wu, University of Minnesota; Martha L. Thurlow, University of Minnesota; Sheryl S. Lazarus, University of Minnesota

State assessment data for 2007-08 to 2013-14 were analyzed to investigate trends in participation, performance, and accommodations received by special education students in grades 3-8 and high school. Results showed differences in participation and performance by content areas and grade level. Trends were also seen for accommodations received.

Grades 1–12 English Learners' Use of Accessibility Features, in Online Language Assessments

Ahyoung Alicia Kim, WIDA, University of Wisconsin-Madison; Meltem Yumsek, University of North Carolina - Greensboro; Mark Chapman, University of Wisconsin - Madison; Howard Gary Cook, University of Wisconsin

This study examines how approximately 1.3 million Grades 1-12 English learners (ELs), including ELs with disabilities, use accessibility features embedded in an online assessment. Findings indicate frequency of ELs' use of eight accessibility features, such as highlighter, and the effect of English proficiency on the use of the features.

Monday, April 8, 2019
4:10 – 6:10pm, Quebec, Paper Session

It's About Time: Considerations of Response Time in Psychometric Models

Discussant: Richard J. Patz, University of California Berkeley

Detecting Carelessly Invalid Responses in Item Batteries Using Item-Level Response Times

Ulf Kroehne, German Institute for International Educational Research; Janine Buchholz, German Institute for International Educational Research (DIPF); Frank Goldhammer, DIPF | Leibniz Institute for Research and Information in Education

Item-level and average responses times extracted with finite state machines from log data of the item batteries used in the PISA 2015 background questionnaire show a bimodal distribution. This response time distribution can be explained by response patterns, allows to identify fast responders and give reasons to filter for straightlining.

Why Does Rapid Guessing Behavior Increase Across Testing Time: Motivational and/or Cognitive Failure?

Marlit Annalena Lindner, IPN - Leibniz Institute for Science and Mathematics Education; Gabriel Nagy, Leibniz Institute for Science and Mathematics Education; Oliver Ludtke, Leibniz Institute for Science and Mathematics Education (IPN), Kiel

We use multivariate latent growth curve modeling to investigate how rapid-guessing-behavior (RGB) is related to the development of cognitive and motivational resources across time. Data indicate that the initial RGB prevalence is a function of both resources, but the RGB increase across time was only related to decreasing motivational resources.

Dealing With Item Nonresponses in Item Response Theory Models Using Item Response Times

Jing Lu, Northeast Normal University; Chun Wang, University of Minnesota; Jian Tao, Northeast Normal University

In this paper, we demonstrate that RT model can be served as missing data model to account for the not-reached responses; we propose a new model with an item-by-person level survival function to explain omitted responses; and we also present a cohesive framework to account for both types of nonresponses.

Response Time and Achievement: Does Spending More Time Increase Test Scores?

Hyo Jeong Shin, Educational Testing Service; Eun Hye Ham, Kongju National University

The present research investigates whether spending more time is likely to bring about increased test scores. Using a matching-based analysis, we compare the test scores between groups of students who spent little time and who spent more time. We illustrate our methods using one of the well-known international large-scale assessments.

An Investigation of Methods to Identify Rapid-Guessing Behavior for a Working-Memory Test

Ping Yin, HumRRO; Mary R. Pommerich, Defense Personnel Assessment Center

This study evaluates two empirical methods for identifying rapid-guessing behavior for a working memory test. The test requires examinees to make rapid, but simple calculations. The study evaluates whether it is possible to identify rapid-guessing behavior for such a test. A modified normative threshold approach was proposed in the study.

Monday, April 8, 2019
4:10 – 6:10pm, Salon A, Coordinated Session

Data Visualization in Standard-Setting

Chair: Carl Setzer, AICPA
Discussant: Richard M. Luecht, University of North Carolina - Greensboro

Standard setting procedures are a critical component of test development. Validity evidence for standard setting usually involves documentation and defensibility of the procedures. However, there has been little to no emphasis on the graphical nature of information exchange between moderators and participants/stakeholders. In many cases, graphics are used to provide calibration and/or impact feedback. The quality of the graphical presentations should be considered, as improper or inefficient usage can result in miscommunication, rather than insight. The purpose of this coordinated session is to emphasize the importance of principled data visualization in standard setting. The first presenter will provide an overview of data visualization principles and where these can be applied in standard setting. Each subsequent presenter will show techniques they have used to present specific information to panelists and/or stakeholders. The discussant will provide perspective on the principles of data visualization in standard setting and give specific feedback regarding the graphic examples.

Do You See What I See? Visual Displays in Standard-Setting
Carl Setzer, AICPA

Visualizing Rater Consistency in a Standard-Setting Process
John T. Willse, University of North Carolina at Greensboro

Real-Time Analytics and Data Visualization During Standard-Setting With Shiny R
Joshua T. Goodman, National Commission on Certification of Physician Assistants

Picture This: Using Visualizations in Standard-Setting for Educator Certification
April L. Zenisky, University of Massachusetts - Amherst; Stephen G. Sireci, University of Massachusetts - Amherst; Maritza Casas, University of Massachusetts - Amherst

Enhancing Panelist Understanding: Visualizations and Standard-Setting
Karla Egan, EdMetric

Monday, April 8, 2019
4:10 – 6:10pm, Salon B, Coordinated Session

Test Adaptation, Translation Errors, and Linking Across Languages in International Large-Scale Assessments

Chair: Steve Dept, cApStAn (Belgium)
Chair: Kadriye Ercikan, Educational Testing Service, Princeton, NJ 08541
Chair: Lale Khorramdel, ETS
Discussant: Ronald K. Hambleton, University of Massachusetts - Amherst

The evaluation of translation procedures, designs and errors in international large-scale assessments such as PISA (Programme for International Student Assessment), PIAAC (Programme for the International Assessment of Adult Competencies) or other educational cross-country surveys, is crucial as resulting test scores need to be comparable across different countries and languages. The proposed symposium will give an overview of state of the art designs and procedures for test adaptation and translation in international large-scale assessments such as PISA, PIAAC and TIMSS. Moreover, it will illustrate the methods and analyses used to anticipate and detect possible errors and to improve the translation in order to achieve valid and comparable international test scores. We will discuss test adaptation and linguistic quality assurance procedures and how translation review procedures can be designed to minimize measurement error caused by translation errors. We will give an overview of research on sources of DIF in translated tests in the past thirty years and we will illustrate how item response theory (IRT) analyses are currently used to test and correct for translation errors to achieve comparable test scores. Moreover, the use of IRT methods to analyze modifications to trend items will be discussed from a linguistic perspective.

Translation, Test Adaptation, and Linguistic Quality Assurance Procedures in PISA and the Program for the International Assessment of Adult Competencies
Steve Dept, cApStAn (Belgium)

Minimizing Measurement Error in PISA due to Translation Errors: Designing Review Procedures
Guillermo Solano-Flores, Stanford University

30 Years of Research on the Sources of Differential Item Functioning in Translated Tests
Avi Allalouf, National Institute for Testing and Evaluation (NITE)

Achieving Comparable PISA and Program for the International Assessment of Adult Competencies Scores Through Linking and Detecting Translation Errors
Frederic Robin, ETS; Lale Khorramdel, ETS

Impact of Trend Block Modifications on Achievement: Lessons Learned in TIMSS 2007–2015
Paulina Korsnakova, IEA

Monday, April 8, 2019
4:10 – 6:10pm, Territories, Paper Session

Innovations in Assessing Student Outcomes

Discussant: Dale Whittington, Retired

Examining Current Practice Regarding Technology-Enhanced Items in K–12 Testing Programs

Sebastian Moncaleano, Boston College; Mike K. Russell, Boston College

This paper examines current practice regarding the use of technology-enhanced items in assessment programs worldwide, with particular focus on utility. Findings suggest that a variety of TEIs are being used by testing programs. Content-specific interactions add measurement value in the form of fidelity, while generic interactions often do not.

Miscarry of Mathematics Learning From Online Games to High-Stakes Assessments

Pamela Paek, ACT; Maria Ofelia San Pedro, ACT, Inc.; Andrew Coulson, MIND Research Institute; Anthony Claypool, Distinctive Schools; Christy Krier, Distinctive Schools

Students' challenges in mathematics result from their surface-level conceptual understanding and sets of flawed processes they cannot justify or explain. This study discusses where students struggled across concepts, understood concepts in one context but not another (games versus high-stakes assessments), and where the disconnects, misconceptions, and misunderstandings may lie.

An Assessment for Introductory Programming Concepts in Middle School Computer Science

Shuchi Grover, Looking Glass Ventures

Teaching of computer science (CS) and programming is rapidly expanding in formal K-12 schooling. However, availability of high-quality assessments for measuring student learning in introductory CS has remained a challenge. This paper discusses the design and features of a summative pencil-paper instrument for assessing introductory programming in middle school CS.

Measuring Proficiency Using Interactive Simulation Data: Empirical Comparison of Evidence Aggregation Methods

Jinnie Choi, Pearson; Kristen E. Dicerbo, Pearson; Matthew Ventura, Pearson Education, Inc.; Emily R. Lai, Pearson; Jim Wood, Minnesota Department of Education; Judi Iverson, Minnesota Department of Education

We empirically compared four methods we can use to aggregate the individual pieces of evidence of learning that can be derived from a learner's interaction with a simulation: item response model, cognitive diagnostic model, Bayesian networks, and percent correct scores. Preliminary results showed positive external validity of all four methods.

Using Bayesian Networks to Characterize Student Performance on Multiple Assessments of Standards

Jiajun Xu, University of Georgia; Nathan Dadey, The National Center for the Improvement of Educational Assessment, Inc.

This paper examines an exploratory way to summarize multiple modular assessments from a large-scale, operational program of interim assessment using Bayesian networks. We follow a data-driven approach to best reflect the empirical relationships between these assessments, and a learning progression approach to provide insight into student learning.

Monday, April 8, 2019
6:30 – 8:30pm, Salons 1 & 2, 19th Floor

We're Here Too: Researchers from Marginalized/Underrepresented Groups Network Reception

Index

A

B

Index

C

Index

D

E

F

Index

G

H

Index

Index

L

Index

M

Index

Index

Q

R

S

Index

T

Index

Index

X

Y

Z

Begin Time	End Time	Room	Floor	Type	Session Title
THURSDAY, APRIL 4, 2019					
8:00am	12:00pm	Algonquin	Mezzanine	Training Session	Introduction to R Software and Applications
8:00am	12:00pm	Salon A	Convention Floor	Training Session	NAEP Response Process Data
8:00am	5:00pm	Salon B	Convention Floor	Training Session	Bayesian Networks in Educational Assessment (Book by Springer)
8:00am	5:00pm	Quebec	Mezzanine	Training Session	Cognitive Diagnosis Modeling: A General Framework Approach and Its Implementation in R
8:00am	5:00pm	Alberta	Mezzanine	Training Session	LNIRT: Joint Modeling of Accuracy and Process Data
8:00am	5:00pm	British Columbia	Mezzanine	Training Session	Measuring Social, Emotional, and Self-Management Skills for Schools and the Workplace
1:00pm	5:00pm	Salon A	Convention Floor	Training Session	Analyzing Features of Assessment Items: An Introduction
1:00pm	5:00pm	Algonquin	Mezzanine	Training Session	Optimal Test Design Approach to Fixed and Adaptive Test Construction Using R
1:00pm	5:00pm	Manitoba	Mezzanine	Training Session	Using R Markdown to Automatically Generate Technical, Research, and Score Reports
FRIDAY, APRIL 5, 2019					
8:00am	12:00pm	Algonquin	Mezzanine	Training Session	A Visual Introduction to Computerized Adaptive Testing
8:00am	12:00pm	Quebec	Mezzanine	Training Session	Computerized Multistage Adaptive Testing: Theory and Applications (Book by Chapman and Hall)
8:00am	12:00pm	Salon B	Convention Floor	Training Session	Nonparametric Cognitive Diagnosis and Computer Adaptive Testing for Small Samples
8:00am	12:00pm	Manitoba	Mezzanine	Training Session	Tips and Tricks to Effectively Communicate Results: Best Practices in Data Visualization
8:00am	5:00pm	Salon A	Convention Floor	Training Session	Exploring, Visualizing, and Modeling Big Data With R
8:00am	5:00pm	Alberta	Mezzanine	Training Session	Generalizability Theory and Applications
8:00am	5:00pm	British Columbia	Mezzanine	Training Session	Learning More From Test Data: New Tools for Test Scoring
1:00pm	5:00pm	Salon B	Convention Floor	Training Session	An Introduction to the Use of Telemetry Data in Video Game Analyses
1:00pm	5:00pm	Confedera-tion 6	Mezzanine	Training Session	Diagnostic Classification Models Part II: Advanced Applications
1:00pm	5:00pm	Territories	Mezzanine	Training Session	Software Packages for Item Response Theory–Based Test Simulation: WinGen3, SimulCAT, MSTGen, and IRTEQ
1:00pm	5:00pm	Quebec	Mezzanine	Training Session	Tools and Strategies for the Design and Evaluation of Score Reports

Begin Time	End Time	Room	Floor	Type	Session Title
1:00pm	5:00pm	Algonquin	Mezzanine	Training Session	Using SAS for Monte Carlo Simulation Studies in Item Response Theory
1:00pm	5:00pm	Manitoba	Mezzanine	Training Session	Vertical Scaling Methodologies, Applications, and Research
3:00pm	4:00pm	Library	Mezzanine	NCME Session	New Board Member Orientation
4:00pm	7:00pm	Library	Mezzanine	NCME Session	Board of Directors Meeting: Opening
4:00pm	6:00pm	Kelly's Landing		NCME Session	Graduate Student Social
SATURDAY, APRIL 6, 2019					
6:30am	7:30am	Tutor 8	Mezzanine	NCME Session	Sunrise Yoga
8:00am	10:00am	Salon A	Convention Floor	Invited Speaker Session	2019 NCME Awards Session
8:00am	10:00am	Algonquin	Mezzanine	Paper Session	Advances in Cognitive Diagnostic Modeling
8:00am	10:00am	Quebec	Mezzanine	Paper Session	Advances in the Evaluation of Item Response Theory Models
8:00am	10:00am	Ballroom	Convention Floor	Coordinated Session	Applications of Multilevel Item Response Theory Models for Collecting Validity Evidence in Educational Assessments
8:00am	10:00am	Salon B	Convention Floor	Paper Session	Emerging Research in Multistage Testing
8:00am	10:00am	Territories	Mezzanine	Paper Session	Pioneering Work in AIG
8:00am	10:00am	Manitoba	Mezzanine	Paper Session	Practical Applications of Validity Research
8:00am	10:00am	British Columbia	Mezzanine	Coordinated Session	Technology-Enhanced Items: Lessons Learned and Future Directions
8:00am	10:00am	Alberta	Mezzanine	Invited Speaker Session	Women in Measurement: Their Unique Contributions
10:25am	11:55am	Ballroom	Convention Floor	Coordinated Session	A Tricky Balance: The Challenges and Opportunities of Balanced Systems of Assessment
10:25am	11:55am	Algonquin	Mezzanine	Coordinated Session	Advancing the Measurement Field With Data Science
10:25am	11:55am	Manitoba	Mezzanine	Coordinated Session	Blending Evidence-Centered Design and Universal Design for Learning in Next-Generation Science Assessment
10:25am	11:55am	Salon A	Convention Floor	Paper Session	CAT: New Directions and Opportunities
10:25am	11:55am	Quebec	Mezzanine	Invited Speaker Session	Communicating Your Research to the Media
10:25am	11:55am	Imperial Room	Main Floor	Electronic Board Session	Electronic Board Session 1
10:25am	11:55am	Salon B	Convention Floor	Paper Session	Emerging Research in Linking and Equating

Begin Time	End Time	Room	Floor	Type	Session Title
10:25am	11:55am	Territories	Mezzanine	Coordinated Session	Practical Measurement for Improvement Science: Principles and Applications
10:25am	11:55am	Alberta	Mezzanine	Coordinated Session	Research on Test-Taking Motivation: Implications for Test Development and Educational Policy
10:25am	11:55am	British Columbia	Mezzanine	Coordinated Session	Scaling Up Assessment Literacy in Teacher Preparation Programs: A Panel Discussion
12:20pm	1:50pm	Ballroom	Convention Floor	Coordinated Session	Automated Assessment of Scientific Reasoning: Developments in the Field
12:20pm	1:50pm	Alberta	Mezzanine	Invited Speaker Session	Classroom Assessment and Educational Measurement
12:20pm	1:50pm	Salon A	Convention Floor	Paper Session	Collecting and Communicating Validity Evidence
12:20pm	1:50pm	Quebec	Mezzanine	Invited Speaker Session	Communicating/Depicting Results in Easily Accessible Ways, Across Broad Audiences
12:20pm	1:50pm	Salon B	Convention Floor	Paper Session	Cultural Considerations in Test Development and Validity
12:20pm	1:50pm	Manitoba	Mezzanine	Paper Session	Fairness Issues in Test Construction
12:20pm	1:50pm	Imperial Room	Main Floor	Graduate Student Poster Session	GSIC Graduate Student Poster Session 1
12:20pm	1:50pm	Algonquin	Mezzanine	Coordinated Session	Scales and Norms for Achievement and Growth: Approaches and Applications
12:20pm	1:50pm	British Columbia	Mezzanine	Paper Session	Research Advancing Item Calibration Methods
12:20pm	1:50pm	Territories	Mezzanine	Coordinated Session	The Assessment of English Writing Skills in Secondary Education in Europe
2:15pm	3:45pm	Ballroom	Convention Floor	Coordinated Session	Advanced Psychometrics for Process Data Analysis in Large-Scale Assessments
2:15pm	3:45pm	Manitoba	Mezzanine	Coordinated Session	Assessment Literacy: What Do They Want to Learn? Four Perspectives
2:15pm	3:45pm	Salon B	Convention Floor	Paper Session	Challenges in Standard Setting
2:15pm	3:45pm	Imperial Room	Main Floor	Electronic Board Session	Electronic Board Session 2
2:15pm	3:45pm	Alberta	Mezzanine	Coordinated Session	Measurement in Adaptive Learning Systems: Challenges and Solutions
2:15pm	3:45pm	British Columbia	Mezzanine	Paper Session	New Insights in Test Assembly
2:15pm	3:45pm	Algonquin	Mezzanine	Coordinated Session	Score Reporting in Ongoing Testing Environments: Reporting Challenges and Innovative Solutions
2:15pm	3:45pm	Quebec	Mezzanine	Coordinated Session	Strengthening the Meaning and Utility of Test Scores for Their Intended Uses

Begin Time	End Time	Room	Floor	Type	Session Title
2:15pm	3:45pm	Territories	Mezzanine	Paper Session	The Role of Student Interest and Engagement on Performance
2:15pm	3:45pm	Salon A	Convention Floor	Coordinated Session	Updating Career Readiness Assessments: Strategies, Challenges, and a Multimethod Validation Approach
4:10pm	6:10pm	Territories	Mezzanine	Invited Speaker Session	Assessment Literacy: Tactics for Traction and Strategies for Success
4:10pm	6:10pm	Ballroom	Convention Floor	Coordinated Session	Automated Scoring Validity Research for a National Large-Scale Writing Assessment
4:10pm	6:10pm	Manitoba	Mezzanine	Coordinated Session	Beyond Learning Progressions: Maps as Assessment Architecture
4:10pm	6:10pm	Salon A	Convention Floor	Paper Session	Communicating Performance Results to Various Audiences
4:10pm	6:10pm	Salon B	Convention Floor	Paper Session	Equating: Applications and Insights
4:10pm	6:10pm	Alberta	Mezzanine	Invited Speaker Session	NCME Session on Excellence in Public Communications
4:10pm	6:10pm	British Columbia	Mezzanine	Paper Session	The Need for Speed? Practical Assessment Implications
4:10pm	6:10pm	Algonquin	Mezzanine	Coordinated Session	Utilizing Expert Judgments to Facilitate Scaling of Tests Adapted for Small Populations
6:30pm	8:00pm	Concert Hall	Convention Floor	NCME Session	NCME and Division D Reception
SUNDAY, APRIL 7, 2019					
8:00am	10:00am	Concert Hall	Convention Floor	NCME Session	NCME Breakfast, Business Meeting, and Presidential Address
10:00am	12:30pm	Manitoba	Mezzanine	NCME Session	NAAD Business Meeting
10:20am	11:50am	Salon A	Convention Floor	Paper Session	Applications of Social and Emotional Learning Measures
10:20am	11:50am	Ballroom	Convention Floor	Coordinated Session	Communicating and Reporting Student Growth
10:20am	11:50am	Alberta	Mezzanine	Coordinated Session	Evaluating Test Speededness in NAEP Digitally Based Assessments
10:20am	11:50am	British Columbia	Mezzanine	Coordinated Session	Measurement and Communication Challenges in a Technology-Based Book Reading Intervention
10:20am	11:50am	Algonquin	Mezzanine	Coordinated Session	New Challenges in Variance Estimation for Digital-Based Educational Assessment
10:20am	11:50am	Territories	Mezzanine	Paper Session	New Directions in Cognitive Diagnostic Modeling
10:20am	11:50am	Salon B	Convention Floor	Paper Session	Technical Considerations in Factor Analysis and Structural Equation Models

Begin Time	End Time	Room	Floor	Type	Session Title
10:20am	11:50am	Quebec	Mezzanine	Coordinated Session	Testing, Testing: Retesting and Inequality in Large-Scale College Admissions Tests (Diversity Issues and Testing Committee's Selected Session)
12:10pm	1:40pm	Alberta	Mezzanine	Coordinated Session	Communicating Achievement Results That Incorporate Response Time Data: Challenges and Advances
12:10pm	1:40pm	Ballroom	Convention Floor	Coordinated Session	Communicating Assessment Results: How to Inform Decision Making in Education
12:10pm	1:40pm	Algonquin	Mezzanine	Coordinated Session	Comparing Automated Scores With Human Scores of Essays in Writing Assessments
12:10pm	1:40pm	British Columbia	Mezzanine	Coordinated Session	Computational Psychometrics for Learning and Assessment in Virtual Environments
12:10pm	1:40pm	Quebec	Mezzanine	Paper Session	Important Considerations in CAT and Item Pool Utilization
12:10pm	1:40pm	Salon A	Convention Floor	Paper Session	Investigating Student Growth and Learning
12:10pm	1:40pm	Salon B	Convention Floor	Paper Session	Measurement and Policies Surrounding Accountability Testing
12:10pm	1:40pm	Territories	Mezzanine	Paper Session	New Directions in Item Response Theory
3:20pm	4:50pm	Alberta	Mezzanine	Invited Speaker Session	2019 NCME Career Award Session
3:20pm	4:50pm	Manitoba	Mezzanine	Paper Session	Advances in Differential Item Functioning Detection and Research
3:20pm	4:50pm	Algonquin	Mezzanine	Coordinated Session	Do Medical Licensing/Certification Exams Really Make a Difference?
3:20pm	4:50pm	Ballroom	Convention Floor	Paper Session	Issues and Advances in Automated Scoring of Constructed Response Items
3:20pm	4:50pm	Salon B	Convention Floor	Paper Session	Lost and Found: Techniques for Handling Missing Data
3:20pm	4:50pm	British Columbia	Mezzanine	Paper Session	Score Comparability: Matters of Mode
3:20pm	4:50pm	Quebec	Mezzanine	Paper Session	Technical Considerations in Item Response Theory
3:20pm	4:50pm	Salon A	Convention Floor	Coordinated Session	The Estimation and Scaling of Rater Effects Parameters for Large-Scale Rater Monitoring
3:20pm	4:50pm	Territories	Mezzanine	Invited Speaker Session	The Influence of Stakeholder Needs and Values on Assessment Design and Reporting
5:05pm	6:35pm	Imperial Room	Main Floor	Electronic Board Session	Electronic Board Session 3
5:05pm	6:35pm	Quebec	Mezzanine	Coordinated Session	Evaluating Teachers' Interpretation and Use of Results in Various Assessment Contexts
5:05pm	6:35pm	Salon A	Convention Floor	Paper Session	New Insights on Engagement, Learning, and Performance
5:05pm	6:35pm	Alberta	Mezzanine	Paper Session	New Learning in Item Analysis Research

Begin Time	End Time	Room	Floor	Type	Session Title
5:05pm	6:35pm	Manitoba	Mezzanine	Coordinated Session	Pioneering a New Approach to Test Design and Development
5:05pm	6:35pm	Territories	Mezzanine	Paper Session	Technical Considerations in Measuring Social and Emotional Learning
5:05pm	6:35pm	Algonquin	Mezzanine	Coordinated Session	Useful and Usable Learning Analytics
5:05pm	6:35pm	Salon B	Convention Floor	Invited Speaker Session	Using the ACT and SAT for Accountability Under the Every Student Succeeds Act: Appropriate or Inappropriate Use
5:05pm	6:35pm	British Columbia	Mezzanine	Coordinated Session	What About Psychometrics in Formative Assessments?
6:45pm	8:15pm	Ballroom	Convention Floor	NCME Session	President's Reception – by invitation only
MONDAY, APRIL 8, 2019					
5:45am	7:00am	Fairmont Lobby		NCME Session	NCME 5k Run/Walk
8:00am	10:00am	Manitoba	Mezzanine	Paper Session	Advances in Evaluating Psychometric Models
8:00am	10:00am	British Columbia	Mezzanine	Paper Session	Advances in Test Security
8:00am	10:00am	Quebec	Mezzanine	Paper Session	Applied Issues in Large-Scale Assessments
8:00am	10:00am	Alberta	Mezzanine	Invited Speaker Session	Appropriately Interpreting, Comparing, and Communicating Results From International Assessments: Challenges and Opportunities
8:00am	10:00am	Salon A	Convention Floor	Coordinated Session	Best Practices Around Automated Scoring Standards
8:00am	10:00am	Territories	Mezzanine	Paper Session	Examining Impacts of Rater Effects
8:00am	10:00am	Algonquin	Mezzanine	Coordinated Session	Optimizing Digital Affordances in NAEP Assessment Tasks: Findings From Two Research Studies
8:00am	10:00am	Ballroom	Convention Floor	Coordinated Session	Raising a New Generation of Measurement Experts: Stories From Around the World
8:00am	10:00am	Salon B	Convention Floor	Paper Session	Technical Considerations in Calculating and Evaluating Reliability
10:25am	11:55am	Salon A	Convention Floor	Coordinated Session	Detecting and Managing Testing Irregularities
10:25am	11:55am	Quebec	Mezzanine	Paper Session	Emerging Research on Longitudinal Diagnostic Classification Models
10:25am	11:55am	Alberta	Mezzanine	Invited Speaker Session	Equity-Centered Design in Assessment: Diversity Issues in Testing Committee Invited Session
10:25am	11:55am	Algonquin	Mezzanine	Coordinated Session	Evidence-Centered Design Extensions: Research in EdTech and Game-Based Learning and Assessment
10:25am	11:55am	Ballroom	Convention Floor	Paper Session	Facilitating the NRC's Assessment Triangle

Begin Time	End Time	Room	Floor	Type	Session Title
10:25am	11:55am	Imperial Room	Main Floor	Graduate Student Poster Session	GSIC Graduate Student Poster Session 2
10:25am	11:55am	Manitoba	Mezzanine	Coordinated Session	Keystroke Logs of Writing Processes in Large-Scale Assessments: Analyses and Applications
10:25am	11:55am	Salon B	Convention Floor	Paper Session	New Insights in Differential Item Functioning Analyses
10:25am	11:55am	British Columbia	Mezzanine	Invited Speaker Session	Preparing Students for College and Careers: Theory, Measurement, and Educational Practice
12:00pm	2:00pm	Library	Mezzanine	NCME Session	Past Presidents' Luncheon
12:20pm	1:50pm	Alberta	Mezzanine	Invited Speaker Session	2018 NCME Career Award Session
12:20pm	1:50pm	Quebec	Mezzanine	Paper Session	Blossoming Research in IRTree Models
12:20pm	1:50pm	Manitoba	Mezzanine	Coordinated Session	Developing Technology-Enhanced Items for Measuring Clinical Judgment in Nursing
12:20pm	1:50pm	Imperial Room	Main Floor	Electronic Board Session	Electronic Board Session 4
12:20pm	1:50pm	Algonquin	Mezzanine	Coordinated Session	Examinations of Practices Used in Human Constructed-Response Rating
12:20pm	1:50pm	Ballroom	Convention Floor	Coordinated Session	Formative Assessment in the Disciplines: Advances in Theory and Practice
12:20pm	1:50pm	Salon A	Convention Floor	Paper Session	Innovative Applications of Machine Learning Techniques
12:20pm	1:50pm	Salon B	Convention Floor	Paper Session	New Directions in Scoring and Classification Accuracy
12:20pm	1:50pm	British Columbia	Mezzanine	Coordinated Session	Testing Strategies, Extended Time Accommodation, and Speededness, Using Process Data in NAEP
2:15pm	3:45pm	Manitoba	Mezzanine	Paper Session	Advances in Item Development, Pretesting, and Selection
2:15pm	3:45pm	British Columbia	Mezzanine	Paper Session	Advances in Multidimensional Item Response Theory
2:15pm	3:45pm	Imperial Room	Main Floor	Electronic Board Session	Electronic Board Session 5
2:15pm	3:45pm	Algonquin	Mezzanine	Coordinated Session	Exploration of Issues With Applying Multistage Testing in NAEP
2:15pm	3:45pm	Ballroom	Convention Floor	Coordinated Session	Formative Multimodal Assessment of Collaborative Problem-Solving Skills
2:15pm	3:45pm	Salon B	Convention Floor	Paper Session	Important Considerations in Setting Cut Scores
2:15pm	3:45pm	Salon A	Convention Floor	Paper Session	Issues, Opportunities, and Challenges With Cognitive Diagnostic Modeling
2:15pm	3:45pm	Quebec	Mezzanine	Paper Session	Issues and Challenges With Adaptive Testing
4:00pm	7:00pm	Library	Mezzanine	NCME Session	Board of Directors Meeting: Closing

Begin Time	End Time	Room	Floor	Type	Session Title
4:10pm	6:10pm	Algonquin	Mezzanine	Paper Session	Assessing Dimensionality: Emerging Research and Technical Considerations
4:10pm	6:10pm	Ballroom	Convention Floor	Paper Session	Assessment as Feedback for Teachers and Students
4:10pm	6:10pm	Manitoba	Mezzanine	Paper Session	Assessment of Special Populations and Subgroups
4:10pm	6:10pm	Alberta	Mezzanine	Paper Session	Challenges, Issues, and Opportunities in Interrater Reliability
4:10pm	6:10pm	Salon A	Convention Floor	Coordinated Session	Data Visualization in Standard Setting
4:10pm	6:10pm	British Columbia	Mezzanine	Paper Session	Important Test Administration and Scoring Considerations
4:10pm	6:10pm	Territories	Mezzanine	Paper Session	Innovations in Assessing Student Outcomes
4:10pm	6:10pm	Quebec	Mezzanine	Paper Session	It's About Time: Considerations of Response Time in Psychometric Models
4:10pm	6:10pm	Salon B	Convention Floor	Coordinated Session	Test Adaptation, Translation Errors, and Linking Across Languages in International Large-Scale Assessments
6:30pm	8:30pm	Salons 1 & 2	19th Floor	NCME Session	We're Here Too: Researchers from Marginalized/Underrepresented Groups Network Reception

RENAISSANCE®

National Council on Measurement in Education is very grateful to the following organizations for their generous financial support of our 2019 Annual Meeting

ACT.

CIAIP. 12£

fii AIR
AMERICAN INSTITUTES FOR RESEARCH"

BUROS
CENTER FOR TESTING

eCollegeBoard

CurricLum Associates·

The Enrollment Management Association
+
Yield Your Best

edCounN
because all students count

(@.
Measuring the Power of Learning:•

Graduate Management

HARVARD

flexMIRT.

GRADUATE SCHOOL OF EDUCATION

imbellus

t:!Y! 1BBQ

fim THE UNIVERSIIY OF IOWA

Advanc ED"

001]
Iowa Testing Programs

measured
:iliiill progress..

LSAC®

@ NBME®

@
Pearson
Smarter

RENAISSANCE.

New Meridian

*

Riverside Insights

SSAT
Elite

ss !l g +

UCLA | CRESST

UMassAmherst
College of Education
Center for Educational Assessment

National Council on Measurement in Education
100 North 20th Street, Suite 400, Philadelphia, PA 19103 (215) 461-6263
http://www.ncme.org/