NCME national council on measurement in education

April 26-30 • San Antonio, TX

San Antonio Marriott Rivercenter

2017 Preliminary Program:

- Schedule-at-a-Glance
- Registration & Housing
- Training Sessions

***INVITATION ONLY EVENTS ARE LISTED SEPARATELY**

Schedule-at-a-Glance

Wednesday, April 26 & Thursday, April 27, 2017
Pre-Conference Training Sessions
(Additional Registration Fee Required)

Thursday, April 27, 2017
4:00 p.m. – 7:00 p.m.
NCME Board of Directors Meeting

Thursday, April 27, 2017
4:30 p.m. – 6:30 p.m. *(Graduate Students only)*
Graduate Student Social
Location TBD

Thursday, April 27, 2017
6:30 p.m.
AERA Opening Plenary Session
Henry B. Gonzalez Convention Center

Friday, April 28, 2017
6:30 a.m. – 7:30 a.m.
NCME Sunrise Yoga
Please join us for the second NCME Sunrise Yoga. We will start promptly at 6:30 a.m. for one hour at the Renaissance. Advance registration required ($10) to reserve your mat. NO EXPERIENCE NECESSARY. Just bring your body and your mind. Namaste.

Friday, April 28, 2017
8:15 a.m. – 6:05 p.m.
Conference Education Program

Friday, April 28, 2017
6:30 p.m. – 8:00 p.m.
NCME and Division D Reception

Saturday, April 29, 2017
8:00 a.m. – 9:00 a.m. *(Tickets Required)*
NCME Breakfast and Business Meeting

Saturday, April 29, 2017
9:00 a.m. – 9:40 a.m.
Presidential Address- Mark Wilson

Saturday, April 29, 2017
10:35 a.m. – 6:05 p.m.
Conference Education Program

Saturday, April 29, 2017
12:35 p.m.
AERA Awards Luncheon
Henry B. Gonzalez Convention Center

Sunday, April 30, 2017
5:45a.m. – 7:00a.m.
(Additional Registration Fee Required)
NCME Fitness Run/Walk
Start your morning with NCME's annual 5k Walk/Run. Meet in the lobby of the Marriott Rivercenter at 5:45 AM. Pickup your bib # and t-shirt at the NCME Information Desk in the hotel prior to race day. Transportation will be provided.

Sunday, April 30, 2017
8:15 a.m. - 6:05 p.m.
Conference Education Program

Sunday, April 30, 2017
4:00 p.m. – 7:00 p.m.
NCME Board of Directors Meeting

***INVITATION ONLY EVENTS ARE
LISTED SEPARATELY**

Registration & Housing

Registration

All NCME attendees must register through AERA. Please visit the link below for information on registration, housing and travel. The headquarters hotel for the NCME Annual Meeting is the Marriott Rivercenter.

TO REGISTER:

- Go to http://www.aera.net
- Click” login” at the top right

  Current & former AERA members:
  Log in with your AERA username and password

  All Others:
  Create a user account by clicking on “Activate” or “Create my Account”

- Once logged in, click on “My AERA” at the top right
- Scroll down to the 2017 AERA Annual Meeting heading and click on “Registration & Housing Now Open! Click here to register”
- Verify your contact information and click “Continue at the bottom of the page.
- Complete the Demographics page and click “Proceed”
- Complete registration
- Click “Checkout Now” in the shopping cart section to enter payment information
- Once you complete your registration, you can make your online hotel reservations by clicking on the” Reservation” button in the upper right side of the registration confirmation summary screen

Once you arrive onsite, please check in at the NCME Information Booth to pick-up your badge. A copy of the final program will be available on the NCME website.

NCME Headquarters Hotel: San Antonio Marriott Rivercenter

Marriott Rivercenter
101 Bowie St
San Antonio, TX 78205
Phone: (210) 223-1000

Check-in and Check-out

- Check-in: 4:00 PM
- Check-out: 12:00 PM
- Express Check-In and Express Checkout
- Video Review Billing, Video Checkout

Parking

- On-site parking, fee: 37 USD daily
- Valet parking, fee: 42 USD daily

Transportation from SAT Airport

This hotel does not provide shuttle service.

- Alternate transportation: Super Shuttle; fee: 18 USD (one way); reservation required
- Estimated taxi fare: 22 USD (one way)

Pre-Conference Training Sessions

Wednesday, April 26 and Thursday, April 27, 2017

The 2017 NCME Pre-Conference Training Sessions will be held at the Marriott Rivercenter on Wednesday, April 26 & Thursday, April 27.

Advance registration for the training sessions is strongly encouraged. The only way to register in advance for these sessions is to use the AERA online registration system. There is a link on the NCME website at http://www.ncme.org/NCME.

Registration onsite will be available only for those training sessions that have not been filled through advance registration

Participants should download the software required prior to the training sessions. Internet connectivity will be available for a few selected training sessions only.

Wednesday, April 26, 2017 Full Day Sessions

8:00 a.m.-5:00 p.m. **Attendees may register for this either as a full day session OR can choose to attend either just the morning (8-12pm) or afternoon (1-5pm) session.**

Bayesian Networks in Educational Assessment

Duanli Yan, Russell G. Almond, Roy Levy, and Diego Zapata-Rivera

The first part of the course will provide background information on Bayesian networks, Graphical Models, and related inference and representation methods and provide examples of their use in educational assessment. The course primarily concentrates on the knowledge engineering processes used to go from a basic conception of a domain to a Bayesian network model which can be used to score a low-stakes assessment. It includes hands-on training in Netica, a commonly used package for Bayesian network analysis.

The second part of the course focuses on building and validating networks using data; covering both EM and MCMC algorithms. It will feature some hands-on work with the R package RNetica, which allows Netica networks to be built from R, and BUGS, which conducts MCMC estimation. It will also include a brief survey of extensions of Bayesian networks; in particular, dynamic Bayesian networks which extend Bayesian networks over multiple points in time.

The previously offered Bayesian network training section has been re-organized into parts which are designed to make sense separately.

Part 1: Bayesian Network Basics (4-hours)

1. Drawing Graphs based on Expert Knowledge
2. Conditional Independence Relationships
3. Example Networks
4. ACED: ECD in Action Demonstration.

Part 2: Building and Validating Bayes Nets from Data (4-hours)

1. Parameterization of Conditional Probability Tables
2. Estimating parameters with the EM algorithm
3. Estimating parameters with MCMC
4. Dynamic Bayesian Networks.

The first part (4-hour block) is meant to be an introduction to Bayesian networks, their properties and their use as a scoring engine. It is a condensation of the original class into 4-hours, focusing on the hands-on experiences. The second part (4-hour block) focuses on building networks from data and recent developments in Bayesian networks. In particular, the second part is designed to be of interest to people who had previously taken the Bayes net course and are now interested more recent developments.

The first part of the course is intended for people who have a good knowledge of probability and statistics (at the level of a college course in statistics with mathematics), but little experience with graphical models (Bayes nets) and related technologies. The second half is designed for people who have some experience with Bayesian networks (perhaps from the earlier part of the course), and who are comfortable with more computationally intensive methods.   In particular, the second part of the course is intended for people who have taken previous versions of this tutorial and want information about new developments in software and algorithms.

8:00 a.m.-5:00 p.m.

Shadow-test Approach to Adaptive Testing

Wim J. van der Linden and Michelle D. Barrett

Until the arrival of PCs and the Internet, adaptive testing was a research topic mainly approached from a purely statistical point of view. However, as soon as testing organization began adopting it, they became overwhelmed with a host of nonstatistical, from a practical perspective much more important issues, such as content blueprint realization, adaptive selection of items organized around common stimuli, prevention of the best items from getting overexposed, the need to select items that fit the same time slot for different examinees equally well, portability of the algorithm across different applications, and the question of how to deal with technical interruptions of the testing process.

The danger of addressing each of these questions and requirements separately is a computer algorithm with a patchwork of heuristic solutions but not much control of their tradeoffs and no guarantee of optimality. For instance, a content blueprint for an adaptive test may be realized by rotating the item selection across its content categories (one of the current solutions) but the price to be paid is loss of precision of the ability scores, overexposure of the most popular items, impossibility to continue item selection within sets, differential time pressure between examinees, etc. Any attempt to fix violations of some of the requirements is bound to lead to new violations of the others. Also, the introduction of each new type of requirement involves recoding of the algorithm.

The shadow-test approach guarantees both satisfaction of all requirements and statistically optimal item selection. In addition, it collects all requirements in a configuration file that can easily be modified when moving from one adaptive testing application to the next. The approach works for any statistical item-selection criterion and is robust against interruptions of testing process because of its permanently updated projection of the remaining part of the test for each examinee.

More technically, the approach is built on the idea of adaptive testing as a sequence of reassembled fixed test forms ("shadow tests") with the most informative free item from each form administered to the examinee. All requirements are collected in a constraint set modeled as a mathematical mixed integer program (MIP) imposed on the reassembled shadow tests using a standard MIP solver. The MIP file serves as the configuration file for the adaptive test; modification of it only involves the addition of a new mathematical inequality or the change of the bound in an existing one. The traditional item-selection criterion is used only to pick the best item from a shadow test rather than from the entire pool; it is not constrained in any way. As the solver is used in a "hot-start mode," using the previous reassembled shadow test as the initial solution for the next, item selection from real-world-size item pools is performed in milliseconds.

The presenters, who have multiple years of experience with the shadow-test approach both through numerous research studies and software developments projects, are currently involved in the development of a cloud-based adaptive testing service that implements the approach.

8:00 a.m.-5:00 p.m.

Cognitive Diagnostic Modeling: A General Framework Approach and its Implementation in R

Jimmy de la Torre and Wenchao Ma

The workshop has a three-fold objective, namely, (1) to provide an overview of CDMs and some recent developments therein, (2) to give participants a hands-on experience conducting various CDM analyses using PR assessment-based data, and (3) to introduce participants to the GDINA as a tool for carrying out comprehensive CDM analyses.

Unlike traditional item response models, CDMs aim to provide information that is finer-grained and more relevant to classroom instruction and learning. As a state-of-the-art methodology, CDM is not typically offered as a regular course in most measurement programs, and may be novel to many practitioners. By giving an overview of CDM, this workshop will be useful to faculty and students specializing in educational measurement, as well as testing professionals working in government or testing organizations.

Furthermore, at present, very few assessments used to provide diagnostic information are developed using a CDM framework. An exception is the proportional reasoning assessment developed by the lead instructor of this workshop, Dr. de la Torre, based on an NSF grant. He will share his experience in developing the assessment, which could be useful for participants interested in developing their own diagnostic assessments. Lastly, very few computer programs that can be used for CDM analyses are currently available, and many of them suffer from various limitations. In this workshop, the GDINA R package developed by the instructors based on a general CDM framework will be introduced. This package overcomes several drawbacks in existing software packages, and offers a set of functions for CDM analyses, such as calibration of various diagnostic models, validation of the Q-matrix, and detection of differential item functioning. After this workshop, participants are expected to be able to conduct various CDM analyses using the GDINA package.

8:00 a.m.-5:00 p.m.

Conceptual Frameworks for Aligning Items to ALDs to Enhance Validity Arguments

Christina Schneider and Steve Ferrara

Participants will

- Analyze alignment and misalignment of items to ALDs from operational tests, thereby identifying the problem regarding the relationship between items and ALDs
- Analyze Range ALDs using construction principles used to write ALDs for many large scale assessment programs and a new ALD conceptual framework
- Write items aligned to the Range ALDs based on item difficulty modeling research
- Practice applying the cognitive task of the ID Matching method, including the benefits of not needing probability judgments and borderline student conceptualizations, making it generalizable from large scale assessments to classroom assessments.

Wednesday, April 26, 2017 Half Day Morning Sessions

8:00 a.m.-12:00 p.m.

Vertical Scaling Methodologies, Applications, and Research

Ye Tong and Dr. Michael J. Kolen

In addition to statistical procedures, successful vertical scaling involves many aspects of testing, including procedures to develop tests, to administer and score tests, and to interpret scores earned on tests. Of course, psychometricians who conduct vertical scaling need to become knowledgeable about all aspects of the process. The prominence of vertical scaling, along with its interdependence on so many aspects of the testing process, also suggests that test developers and all other testing professionals 2 should be familiar with the concepts, statistical procedures, and practical issues associated with vertical scaling.

We would like to provide an opportunity for interested individuals to enhance their knowledge to conduct vertical scaling, as well as to fully appreciate the practical consequences of various changes in testing procedures on vertical scaling, such as the consequences of many test-legislation initiatives, the use of performance assessments, and the introduction of computerized test administration.

This training session on vertical scaling considers both the most frequently used linking designs and methodologies and some of the practical issues involved. In addition, how vertical scaling can be used with the common core state standards, with the assessments by the consortia based on the co-directors' perspectives, as well as what types of interpretations can be made of measures of growth towards college readiness, will be discussed.

After attending this pre-session, the participant should be able to:
1. Understand the purposes of vertical scaling and the context in which it is conducted.
2. Distinguish between vertical scaling and other related methodologies and procedures.
3. Understand the distinction between vertical scaling data collection designs and linking methods.
4. Understand the fundamental concepts of vertical scaling – including data collection designs, linking methods, and statistical assumptions necessary for vertical scaling.
5. Compute linking functions and choose among methods.

6. Interpret results from vertical scaling analyses.
7. Design reasonable and useful vertical scaling studies.
8. Conduct vertical scaling in realistic testing situations.
9. Identify appropriate and inappropriate uses and interpretations of vertical scaling results.
10. Understand the research literature on vertical scaling.

The targeted audience is upper-level graduate students and new Ph.D.'s with interest in learning about vertical scaling methodology and practice. In addition, testing professionals with operational or oversight responsibility for vertical scaling, and others with interest in learning about vertical scaling methods and practices could likely benefit from this session. As minimal prerequisites for attendance, participants should have taken two graduate course in measurement and at least two graduate courses in statistics.

This training session has been offered at NCME every two years since 2007 and has been very well received in the past. Given the current interest in vertical scaling, it seems likely that there would be much interest in this session in the 2015 NCME conference as well. For the 2015 session, the co-directors plan to introduce more hands- on examples as well as in-depth discussion within the common core state standards context and the use of measures of growth.

8:00 a.m.-12:00 p.m.

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

Heidi Andrade

A rubric is a useful and versatile tool for classroom assessment. Unfortunately, many rubrics used by teachers are of low quality, not appropriate for use outside a standardized testing context, and/or incorrectly scored. In addition, rubrics are too infrequently used to scaffold the kinds of formative assessment that have been shown to promote learning – but only when the rubrics are of high quality. As a result, rubrics can distort rather than clarify learning targets, provide inaccurate information about student learning, and lie dormant as teaching and learning tools. This half-day, interactive workshop will reveal and resolve common pitfalls in rubric use, drawing on examples from actual classrooms.

The goal of the session is for participants to understand:

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

8:00 a.m.-12:00 p.m.

An introduction to Linking and Equating in R

Anthony Albano and Jonathan Weeks

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

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

A background in introductory statistics and experience using R are recommended but not required. Participants should bring their own computers, with R (R Core Team, 2016) and the most recent versions of the *equate* package (Albano, in press) and *plink* package (Weeks, 2012) installed. Electronic training materials will be provided via email at least one week prior to the conference.

The presenters for this training session have authored two popular R packages for linking and equating. Both have also presented linking and equating workshops at NCME in the past. Anthony Albano is faculty at the University of Nebraska-Lincoln, where he teaches courses in measurement and equating. He leads multiple research projects addressing, for example, issues in presmoothing, equating with nonequivalent groups, small samples, and equating in the context of growth modeling. He also consults with different organizations on practical issues in equating. His qualifications as presenter are based on teaching and advising graduate students, conducting research, programming, and consulting on current equating issues. Jonathan Weeks is an Associate Research Scientist at ETS. He has authored and co-authored a number of papers on unidimensional and multidimensional vertical scaling and has also served as a consultant on several projects requiring the development of both horizontal and vertical scales.

Wednesday, April 26, 2017 Half Day Afternoon Sessions

1:00 p.m.-5:00 p.m.

The History of Educational Measurement in America: Origins to 1950

Michael B Bunch, Michael Beck, Brian Clauser, and Michelle Croft

This half-day training session provides an overview of the historical foundations of educational measurement in the United States, from early influences to about 1950. Its purpose is to provide historical perspective to the current theories, practices, and policies associated with educational measurement.

Module 1 addresses the events of the mid and late 19th Century that led to calls for standardized testing of public school students. From Horace Mann’s calls in 1845 for standardized written tests as a means for evaluating students and their teachers, through the development of the “new type” objective (aka multiple choice) tests of the early 1920s, to the standardization and mechanization of test administration necessitated by increasing demands for accountability in the 1930s and 40s, Mike Beck will show how the solutions to problems of more than 100 years ago still influence our practice today.

Module 2 addresses the social/political/legal issues of the same period, with particular emphasis on the rise of state-operated evaluation of local school districts. During this same period, as reliance on standardized testing increased, opposition also increased. Current controversies over federal control, opting out, and use of student assessments for accountability have deep roots in debates of the 1920s and 1930s. Michelle Croft will show how those past debates and their social/political/legal dimensions are still relevant and what we can learn from them as we address these and similar issues today.

Module 3 addresses the development of theory. Standardization of practice gave rise to a fledgling science of educational measurement, marked by reliance on statistical measures, most notably the correlation coefficient. Brian Clauser will show how this body of knowledge evolved, ultimately leading to the fully developed test theory reflected in Gulliksen's (1950) *Theory of Mental Test Scores*. Along the way, he will show how different theories arose, came into conflict, and resolved.

Module 4 addresses the rise of a professional class of educational measurement specialists from the 1930s through the end of World War II. It focuses on the rise of educational testing organizations – public and private – and their influence on the theory, practice, and public policy of educational measurement. Michael Bunch will trace the development of the profession of educational measurement, with particular emphasis on the contributions of James McKeen Cattell (Psychological Corporation), the Clarks (California Test Bureau, aka CTB), and Carl Brigham (College Board).

This session stops just short of the publication of the first edition of *Educational Measurement*. However, it does provide a preview of that seminal text as well as other key events of the next 65 years, which will be the subject of another pre-session in 2018.

The format of the session is primarily lecture with multiple opportunities for questions and answers at the end of each module and during a question and answer session. Any questions not answered in full during the session will be fully researched following the session, and participants will receive a follow-up set of all questions and answers one month after the conclusion of the session.

1:00 p.m.-5:00 p.m.

Analyzing NAEP Data Using Plausible Values and Marginal Estimation with AM

Emmanuel Sikali and Young Yee Kim

The National Assessment of Educational Progress (NAEP) is the only nationally representative and continuing assessment of what US students know and can do in various subject areas. Assessments are conducted periodically in mathematics, reading, science, writing, the arts, civics, and other subjects. Since NAEP results serve as a common metric for all states and select urban districts, many researchers are interested in conducting studies using NAEP data. However, NAEP data pose many challenges for researchers due to its special design features.

In NAEP, students take different but overlapping combinations of portions of the entire pool of items. No one student receives enough test questions to provide an accurate test score. Thus, NAEP does not provide scores for individual student. Instead, multiple plausible values are provided for each student for secondary data analysts' research.

The unique psychometric and sampling features of NAEP require special considerations in analyzing NAEP data, prohibiting researchers from using common statistical software packages (e.g. SAS) without appropriate handling these considerations. There are two ways of analyzing NAEP data: (1) analysis with plausible values and (2) the marginal estimation approach with item response data.

This workshop will introduce participants to the psychometric design of NAEP, sampling procedures and data analysis strategies required by these design features. These include the marginal maximum likelihood estimation approach to computing scale scores, and appropriate variance estimation procedures. In addition, participants will learn how these two analytic approaches differ, when each approach might be preferable, and how to actually conduct analysis with both approaches. A mini-sample public-use NAEP data file released in 2011 by National Center for Education Statistics (NCES) will be used to learn how to analyze NAEP data. A free software program AM, developed for analysis of NAEP-like data will be used for the class.

At the end of the session, participants are expected to:

- be familiar with the psychometric and sampling design, content, and research utility of the NAEP assessments;
- understand the need for using weighting and variance estimation variables correctly; and
- be able to analyze complex NAEP data either using the plausible values approach or the marginal estimation approach, using the AM software; and
- know the resources available to them at the NCES that can assist their research

1:00 p.m.-5:00 p.m.

Landing Your Dream Job for Graduate Students

Deborah J Harris and Xin Li

This training session will address practical topics graduate students in measurement are interested in regarding finding a job and starting a career, concentrating on what to do now while they are still in school to best prepare for a job (including finding a dissertation topic, selecting a committee, maximizing experiences while still a student with networking, internships, and volunteering, and providing suggestions to the questions regarding what types of coursework an employer looks for, and what would make a good job talk), how to locate, interview for, and obtain a job (including how to find where jobs are, how to apply for jobs --targeting cover letters, references, and resumes), what to expect in the interview process (including job talks, questions to ask, and negotiating an offer), and what's next after they have started their first post PhD job (including adjusting to the environment, establishing a career path, publishing, finding mentors, balancing work and life, and becoming active in the profession).

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

The presenters have been working in the profession for a few years to several decades. They have diverse experiences to draw on. In addition, materials from previous presenters and others interested in fostering graduate student careers are also included. Deborah J. Harris is Vice President of Measurement Research at ACT, Inc., and has been involved with graduate students through teaching as an adjunct at the University of Iowa, serving on dissertation committees, working with graduate Research Assistants, and working as a mentor and an organizer of the ACT Summer Intern Program. She is involved in the hiring and mentoring of staff, which frequently includes new doctorate recipients. Xin Li is a Psychometrician at ACT, and a recent graduate from the University of Texas-Austin. She has had a variety of experiences as a graduate student, and as a relatively new professional, she can speak to finishing up her course work and dissertation, navigating the job search process, and settling into beginning a career. In 2015, we had a human resource (HR) person also attend and provide information related to industry standards in the hiring process, including the current online application systems most organizations use. (The HR person who would be available in 2017 has not yet been identified and therefore is not listed as a presenter.) Previous versions of this training session were presented in 1998, 2000-2002, and in 2005-2016. The session has been well-received by the attendees based on the feedback they have provided.

1:00 p.m.-5:00 p.m.

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

Caroline Wylie and Christine Lyon

Given the frequent call for greater assessment literacy among all educational stakeholders and current pressure to make inferences about assessment data beyond what the data can legitimately support, we believe this topic will be relevant to a wide range of participants. There are calls to reduce testing, while at the same time increase the amount of information provided. Consideration of assessment cycles and grain-size (Wiliam, 2006; Shavelson et al., 2008) can support appropriate interpretation and use of different types of data. Different types of data can be used at across the year for informed decision making (e.g., using summative assessment formatively for curriculum review and unit planning, using interim assessment data to confirm or adjust long range planning, using formative assessment for short-cycle adjustments). The session learning goals are for participants to understand:

- Appropriate uses of summative, interim and formative assessment information
- How to support the development of strategies for the appropriate use of assessment data by schools and districts
- How to create more robust PLC discussions around the use of assessment data

The session will be jointly presented by Caroline Wylie and Christy Lyon. Caroline is a Research Director and Senior Research Scientist at ETS, and focuses her research on applications of formative assessment in a variety of classroom contexts. Caroline was the lead researcher for a formative assessment professional development program created at ETS, and has been the PI/co-PI on multiple IES and NSF studies focused on formative assessment in mathematics and science.

Christine is a lead Research Project Manager at ETS. She was co-developer for a program focused on increasing teachers' formative assessment practice and led research on the characteristics of effective and scalable professional development. More recently, she conducted research to understand and deepen teachers' implementation of *Cognitively Based Assessment of, for, and as Learning (CBAL)* classroom materials. Both presenters have extensive experience engaging teachers, district and state-level staff in assessment conversations.

1:00 p.m.-5:00 p.m.

An Introduction to R for Quantitative Methods

Brian Habing and Jessalyn Smith

This session is designed to introduce the statistical package R so that the attendees will both be able to use R for basic statistical analyses and have an understanding of how it can be used in their own work. This will include guidance on introducing R to students or colleagues, constructing templates that can be adjusted for repeated analyses or classroom lectures, selecting appropriate packages for carrying out more advanced methods, and providing selected custom designed functions to easily produce output in a format that is most helpful.

The training session is organized in three sections: an introduction to R, implementing the standard methods in R, and demonstrations of common (or requested) quantitative procedures with R with a look at some common psychometric packages and their built in functions. Each of the three sections will be accompanied by an html document designed so that the participants can work along with the instructors on their own laptops (either by cutting and pasting code in or by modifying an existing example).

The material in the first third of the course (introducing R) will be presented as it could be to an introductory methods course. An html document provided to each participant will contain a written copy of the explanatory material as well as the code being explained. As the material is covered the participants will first cut and paste in the example code on their own laptops to observe the output, and will then be led through some modifications of the basic examples. The base material for the second portion of the course (standard quantitative methods) will use web-templates designed to parallel to common hypothetical examples used in practice or a text book. For several of the methods participants will verify that the provided code does produce answers similar to that in the provided examples, and modify the code to answer provided exercises. This will include using built in functions, functions from packages, and several custom made functions. The final portion of the course will cover constructing basic simulations, writing functions for a couple common quantitative procedures, and exploring the functions that are currently in some of the most widely used psychometric packages.

Several example datasets will be provided for participants to work with and use to run analyses. Throughout the training session one of the two presenters will be available to help answer both technical and substantive questions as they occur.

Thursday, April 27, 2017 Full Day Sessions

8:00 a.m.-5:00 p.m.

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

Hong Jiao, Yong Luo , and Kaiwen Man

This training session focuses on Bayesian estimation of standard and extended item response theory (IRT) model parameters using two software programs, OpenBUGS and Stan. The intended audience is intermediate and advanced graduate students, researchers, and practitioners who are interested in learning the basics and advanced topics related to IRT model parameter estimation using these two Bayesian estimation software programs. This full-day session introduces the basics of the software program OpenBUGS and Stan. The estimation of model parameters for unidimensional dichotomous and polytomous IRT models, multidimensional including testlet models, and multilevel IRT models will be illustrated and demonstrated.

Learning will occur through lecture, demonstration, and hands-on activities running OpenBUGS and Stan. It is expected the audience will have some basic knowledge of the Bayesian theory, but not required. Attendees will bring their own laptop and download the software programs free online. It is expected that attendees can develop OpenBUGS and Stan codes for new extended IRT models for their own research and psychometric modeling after they master the basics of writing OpenBUGS and Stan codes for standard and extended IRT models.

The software programs for IRT model parameter estimation usually contain limited capacity. Most often only the author(s) of the program has access to the source codes. Thus it is not possible for practitioners and researchers to revise or modify the program to meet their own research or psychometric analysis needs. OpenBUGS has been extensively used in psychometric modeling while Stan is a relatively new Bayesian inference program with higher efficiency. This training session will provide audience with capacity and flexibility to build their OpenBUGS or Stan codes for parameter estimation for more complex or new IRT models that would otherwise be more difficult to attain with commercial software programs.

8:00 a.m.-5:00 p.m.

Diagnostic Classification Models: Theory, Methods, and Applications

Laine Bradshaw and Matthew J. Madison

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

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

8:00 a.m.-5:00 p.m.

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

Patrick C. Kyllonen and Jonas Bertling

The goal of the workshop is to provide participants with an overview of the key character constructs associated with success in education and life, illustrate ways to measure them, and provide hands-on experiences to enable participants developing or choosing their own assessments. The workshop begins with an overview of the issues involved in assessing character skills. We identify a number of frameworks and discuss overlaps between different using communities, such as personality, developmental, and counseling psychology, student affairs and institutional research, K-12 and higher education, and so forth. We examine the use of certain methods in detail, including anchoring vignettes, forced-choice methods, and situational judgment tests. For anchoring vignettes we review writing, scoring, and reporting and include hands-on experience for both. For forced-choice methods we review data collection and different scoring approaches, and their advantages and disadvantages. For situational judgment testing, we review item development, including approaches for collecting critical incident data from experts, using various response formats, and scoring. For all methods we provide hands-on experience in item development and scoring.

8:00 a.m.-5:00 p.m.

An Introduction to Hierarchical Rater Models for the Analysis of Ratings

Jodi M Casabianca, Brian Junker, and Ricardo Nieto

The measurement of individuals using constructed responses is sensitive to rater effects. Both human and machine raters have the potential to introduce error to scores, thereby reducing their precision and impacting the uses of those scores. Several different types of rater effects may be exist in ratings, but most commonly the literature discusses rater bias, a rater's tendency to score higher (leniency) or lower (severity) on average (Kingsbury, 1922). The evaluation of rater behavior and accounting for rater effects in scoring are two important goals in psychometrics. IRT models that include parameters for raters provide a mechanism within which to achieve these goals. While there are several IRT models for ratings data that incorporate parameters to account for, and study, raters and the rating process, only a handful account for the dependencies brought about by multiple ratings of the same work. These models include the hierarchical rater model (HRM; Casabianca, Junker, & Patz, 2016; Patz, Junker, Johnson, & Mariano, 2002), Verhelst and Verstralen's (2001) IRT model for multiple raters, Wilson and Hoskins (2001) rater bundle models and more recently, DeCarlo, Kim, and Johnson's variant of the HRM (2011). It is the hierarchical structure of the HRM that accounts for the dependencies brought about by multiple ratings of the same work thereby avoiding the downward bias in the standard errors of the latent traits (Patz, et al, 2002). Extensions of the basic parameterization share this desirable feature. For longitudinal assessments and research using ratings, the longitudinal HRM (L-HRM) may include individual and overall growth terms and/or a time series component (e.g. an autoregressive model). For assessments using constructed response items based on a multidimensional rubric, the multidimensional HRM (M-HRM) employs the multidimensional generalized partial credit model and permits an analysis of dimensionspecific rater behavior.

The goals are to: (i) teach participants about the utility of the HRM and its extensions, (ii) provide participants the opportunity to practice fitting the HRM, and (iii) provide the tools necessary for participants to successfully fit these models independently. We hope that the training session generates interest in the HRM, thereby leading to additional research and application of the model by others in the field.

8:00 a.m.-5:00 p.m.

A Framework and Platform for the Development of Assessment Literacy

Damian Betebenner, Charles Depascale, Luciana Conchado, Amy Sharpe, and Kelli Ryan

In this full-day session, participants interested in sharing their expertise via the development of assessment literacy modules will be introduced to free, open source tools and a simple workflow that enables them to quickly build interactive assessment modules that are immediately publishable via the web and suitable for collaboration/sharing with others.

The tools and ideas shared in this hack-a-thon session derive from work that the National Center for the Improvement of Educational Assessment (the Center) conducts with clients. In our efforts to build capacity of state and district users and producers of assessment data, our goal is to deliver content to our clients that they can both learn from and further share with their stakeholders to enhance understanding of assessments and their defensible uses. The ubiquitous nature of the web has provided an ideal distribution mechanism for these efforts. In 2015, the Center began development of a GitHub based mechanism that allows users to produce assessment literacy modules that are immediately viewable and shareable by others. The goal in this effort is to produce a platform that enables anyone with the desire and content area expertise and to produce interactive assessment literacy modules.

Our effort to produce this platform was inspired by the work of Bret Victor and his Explorable Explanations http://explorableexplanations.com/. We find current best practices/efforts in assessment literacy are often lacking because of the distance, or barrier, between the materials and the user: They either revert to standard “texts” to explain the issue/concept or, more recently, produce “talking PowerPoints” derived from those texts. Following Victor’s work, we believe that Title: A Framework and Platform for the Development of Assessment Literacy authentic understanding on the part of the user requires going beyond just telling or showing them to actually involving them in the shared content. The rich interactivity of web-based materials and the increasing availability of open source tools and resources allows for the development of this content an achievable reality. Our efforts are directed toward making the barrier to entry in producing such content as low as possible.

Efforts to conceptualize and develop tools for assessment literacy modules were initiated and refined with the participation of summer interns at the Center. The interns were tasked with developing modules and stress testing the tools and workflow used in developing these modules so that the tools simultaneously support (i) the rich content of explorable explanations and (ii) present as low a barrier to entry for new assessment literacy content developers as possible. As we work with clients throughout this year, we will continue to develop modules and refine the open source platform. NCME participants in this workshop/hacka-thon will have access to the modules and a platform that they can immediately leverage to produce high fidelity, assessment literacy content for their own use.

The backbone of the open-source tool set used for the assessment literacy module development hack-a-thon is the freely available GitHub version control social coding website. Version control software is fundamental in any software development project. Version control allows for teams to work on content together in a well-coordinated and structured fashion. GitHub has made version control social by building a web-based version of the tool that both makes Git accessible but, more importantly, allows third parties to build tools and services on top of their tool. To that end, we have built a free/open-source viewer that takes the GitHub hosted content and displays it as a web-based interactive module1.

Thursday, April 27, 2017 Half Day Morning Sessions

8:00 a.m.-12:00 p.m.

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

Jiangang Hao, Alina von Davier, Kristen DiCerbo, and Robert Mislevy

Game/simulation based assessment has a number of advantages over traditional assessment, and is widely considered as an important future direction for assessments (Cope & Kalantzis, 2015). Evidence Centered Design (ECD, Mislevy, Steinberg, & Almond, 2003) provides a theoretical framework for designing game/simulation-based assessments around a validity argument that connects a test taker’s activities to performance on some predefined construct of interest. However, the implementation of ECD principles during the actual development of a game/simulation-based assessment is not necessarily simple. It generally involves rounds of iterations among specialists with different areas of expertise, such as learning scientists, software developers, data scientists, and psychometricians. In particular, assessing a test taker's performance depends on having a well-structured record of the player's activities and situational variables, and having techniques and tools to effectively analyze those records to determine the degree to which the player's activities reveal the targeted constructs.

The process data that record a test taker's activities and the situational/environmental variables are very complex. The skills needed to parse, aggregate and model this type of data are generally not well covered in most educational measurement programs. The goal of this training session is to promote consensus among researchers about how to efficiently implement ECD in practice, summarize the computational psychometrics modeling techniques of the process data, and provide a data solution for efficient and cost-effective analytics for game/simulation-based assessment via a set of hands-on exercises.

The whole session is divided into three parts: a) ECD and its implementation; b) data model, analytics and python programming language; and c) computational psychometric modeling of process data. Part a) is intended to give a practical guide about how the ECD principles are implemented in real game/simulation-based assessments with emphasis on those practical considerations that can make the process efficient. Part b) introduces a comprehensive data solution that involves a set of recommended operational procedures to ensure evidence capturing, a data model for log file structure (Hao, Mislevy, von Davier, & Smith, 2014), and a Python library/package, glassPy (Hao, Smith, Mislevy, von Davier, & Bauer, 2016), for evidence identification, aggregation and analytics. The part c) focuses on summarizing the methodologies used to model the process data generated from game/simulation-based assessments, and some exemplar use cases will be introduced.

8:00 a.m.-12:00 p.m.

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

Mary Veazey and Ye Tong

The move to online assessments offers interesting possibilities for how students and teachers interact with content. This move also requires that assessment developers consider the ramifications of this move and provide a solution appropriate for all stakeholders. In particular, psychometricians and content experts need to understand the various challenges when working to develop a valid assessment that transitions from previous assessments while incorporating the technology that best assesses the content. For any assessment, but especially one incorporating new technology, the issues teachers and students face in the classroom must also be considered.

This training session is an opportunity for anyone interested in the various aspects of online assessment to learn more about the process, better understand the points of view, and gain experience in the unique steps needed to incorporate this technology. The overall goal is that these attendees will be able to appreciate the challenges each group faces. That understanding will better inform decisions made around assessments, leading to more accurate and meaningful data for teachers and students. During the training session, psychometric discussion will center on reliability and validity evidence associated with new online item types, scaling and equating, computeradaptive testing, and mode and device comparability. Selecting item types and working with technology-enhanced items will be the focus on the content discussion. Scoring adjudication will be looked at from both content and psychometrics for common online item types. The classroom perspective on accessibility and authenticity will be shared as well.

In addition to understanding the different points of view on online assessments, attendees should be able to:

- Understand the purpose of moving to online assessments
- Identify and discuss the pros and cons of various item types in an online environment
- Understand when paper-equivalent items for technology-enhanced items may be needed and considerations for developing the items
- Identify and discuss how scoring decisions for online items affect scoring of paper items and vice versa
- Understand the need for adjudication of online item types and its associated logistical challenges
- Discuss with stakeholders the reasons for various assessment decisions

The audience for this training session will be stakeholders who are transitioning from paper to online assessments, especially psychometricians, content developers, educators, and administrators making large-scale assessment decisions. Graduate students and program directors will also benefit from the training, as it will give them insight into the requirements in developing online assessments. Attendees should have some exposure to online environments and functionalities.

The session will focus on the most common item types and situations so that it is applicable across various platforms. Hands-on examples and work samples will be included, along with in-depth discussion with an emphasis on practical challenges and implications. Additional topics about how technology affects assessments, such as artificial intelligence scoring, will also be addressed.

Thursday, April 27, 2017 Half Day Afternoon Sessions

1:00 p.m.-5:00 p.m.

Computerized Multistage Adaptive Testing: Theory and Applications

Duanli Yan, Alina von Davier, and Kyung Han

This course is intended for people who have some basic understanding of item response theory and CAT.

The focus of the workshop will be on MST theory and applications including alternative scoring and estimation methods, classification tests, routing and scoring, linking, test security, as well as a live demonstration of MST software MSTGen (Han, 2013). This workshop is based on the edited volume of Yan, von Davier, & Lewis (2014). The volume is structured to take the reader through all the operational aspects of the test, from the design to the post-administration analyzes. In particular, the chapters of Yan, Lewis, and von Davier; Lewis and Smith; Lee, Lewis, and von Davier; Haberman and von Davier; and Han and Kosinski are the basis for this workshop.

MSTGen (Han, 2013), a computer software tool for MST simulation studies, will be introduced by Han. MSTGen supports both conventional MST by routing mode and the new MST by shaping mode, and examples of both MST modes will be covered. The software is offered at no cost, and participants are encouraged to bring their own computers for a brief hands-on training.

There are four lecture sessions with hands-on examples using MST Gen software. This gives about 4 hours in total. The training course consists of a series of lectures and hands-on examples in the following four sessions:

1. *MST Overview, Design, and Assembly*
2. *MST Routing, Scoring, and Estimations*

3. *MST Applications*
4. *MST Simulation Software*.

1:00 p.m.-5:00 p.m.

Evaluating Alignment of Computer Adaptive Assessments

Katerina Schenke and Deborah La Torre

The introduction of the Common Core has ushered in a comprehensive and more rigorous set of standards on which students across all states should be tested. Because of this, assessments have gone from mostly fixed, paper-pencil forms, to computer-based adaptive assessments (CAT) that claim to be more reliable and efficient in estimating student proficiency. Part of this shift requires test administrators to evaluate the test delivery system in new and different ways. Specifically, the appropriateness of the inferences drawn from test scores for a given purpose or test score use requires more comprehensive evidence-based arguments for validity of computer adaptive standardized tests. Current thinking in item alignment draws on work from Wise, Kingsbury, & Webb (2015), which focuses specifically on content alignment between the number of items administered on the test and the number of items aligned to a specific content area. What is missing from Wise et al., (2015) is a broader consideration of evaluating item alignment in CATs, which includes not only evaluation of the test instances, but also an evaluation of the items themselves. The purpose of this workshop is to provide theoretical and hands-on practical support in evaluating item-alignment in CAT environments.

A substantial match between items and test instances would provide evidence that inferences drawn from test scores could be representative of the content and cognitive demands that the item development team intended. CAT algorithm evaluation can provide evidence that the tests students are administered with adequate representations of the content as specified by states (fidelity of blueprint, or test specifications) and have adequate psychometric properties. A strong match would provide evidence that inferences drawn from test scores could be representative of the intended blueprint, or test specifications, and thus sufficiently representative of a state's intended standards and cognitive complexity for that grade and domain. Information about item difficulty, item exposure, and item information from investigations of the psychometric properties of each administered test form produced by the CAT algorithm is evidence which could be used to build an argument that the CAT algorithm is fairly generating test forms for students at every level of overall achievement.

1:00 p.m.-5:00 p.m.

Using Visual Displays to Inform Assessment Development and Validation

Brett P Foley

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

1:00 p.m.-5:00 p.m.

A Visual Introduction to Computerized Adaptive Testing

Yuehmei Chien

Adaptive tests often provide distinct advantages over fixed linear tests. By personalizing the test for each student, they provide better measurement precision with a shorter test length. The test taker does not waste time on very easy items or struggle with very hard items. Higher measurement precision results can be attained because the items given to each test taker are tailored to their ability and therefore items with higher item information are administered.

Adaptive testing has become more and more prevalent over time. Younger generations of students are getting more comfortable with computerbased learning and testing, having grown up with computers and tablets from an early age. As these testing techniques become more commonplace, it will be important for practitioners and researchers to have deeper understanding of how online adaptive testing works, its advantages and disadvantages, and how to maximize the benefit to themselves and the test takers.

This training session will accomplish several goals. First, the training will provide participants with essential background information on CAT including major CAT components and some consideration for operational CAT setting. Several CAT examples of formative assessment, summative assessment and licensure examines will also be presented. Second, the participants will access online CAT demo applications developed by the presenter, and learn different algorithm by doing some handson CAT configuration activities. Third, a lecture on multistage CAT and automated test assembly (ATA) will be given, which includes the concept of multistage and an introduction of a linear solver that is used for ATA. Lastly, a web application for ATA will be demonstrated and the participants will use it to build several panels for multistage CAT and run a multistage CAT simulation as well.

The presenter is a senior research scientist, a software developer, and a psychometrician with Pearson since 2006. She had designed CAT algorithms and implemented CAT delivery engine and simulator for several testing programs including College Board’s Accuplacer placement tests and Pearson’s K12 assessment CAT programs. She also conducted many CAT related research in the past 10 years. Some of the research topics include variable length adaptive diagnostic testing, custommade CAT, weighted penalty model for content balancing in CAT, and dealing with variability within Item clones in CAT. In recent years, the presenter has been developing psychometric tools with visualization, including an online automated test assembly system and an adaptive diagnostic assessment simulator, which present the results visually and provides insightful information. In the same fashion, the presenter will develop lecture material with interactive visualization in order for the participants to learn CAT concept a step further by interacting with the CAT application and by creating stage panels for multistage CAT.