NCME Annual Meeting Training Sessions

Montreal, Quebec, Canada, CA, April 2005

Admission to training sessions is limited to ticket holders. Advanced registration for training sessions will be offered on-line at the NCME Website. In-person registration at the conference will also be possible, provided space in the workshops is still available. In-person registration will be located at the NCME registration area in one of the convention hotels. Courses are subject to cancellation for insufficient registration. Detailed abstracts of these sessions are provided below.

Please note that the NCME registration link is not yet active.

Please visit the AERA Website (http://www.aera.net/) for training session and conference registration information.

NCME Annual Meeting Training Sessions

Montreal, Quebec, Canada, CA, April 2005

Admission to training sessions is limited to ticket holders. Advanced registration for training sessions will be offered on-line at the NCME Website. In-person registration at the conference will also be possible, provided space in the workshops is still available. In-person registration will be located at the NCME registration area in one of the convention hotels. Courses are subject to cancellation for insufficient registration. Detailed abstracts of these sessions are provided below.

Session 1: The Kernel Method of Observed Score Test Equating
Presenters: *Paul W. Holland, ETS; Alina A. von Davier, ETS; Ning Han, ETS*

Sunday, April 10, 8:30–5:00 Fee: $95

Test equating methods are used to produce scores that are comparable across different test forms. The Kernel Method of Test Equating (KE) is a unified approach to test equating based on a flexible family of equipercentile-like equating functions that contains the linear equating function as a special case. We view any observed–score test equating as having five steps or parts, each of which involves distinct ideas. They are: (1) pre-smoothing; (2) estimation of the score probabilities on the target population; (3) continuization; (4) computing of the equating function; (5) computing the standard error of equating and related accuracy measures. KE brings together these steps into an organized whole rather than treating them as disparate problems. KE exploits pre-smoothing by fitting log-linear models to score data, and incorporates the results of pre-smoothing into step 5 as well. KE provides new tools for comparing two or more equating functions and to rationally choose between them. We will cover the theory behind KE as well as its application to the Equivalent Groups, Single Group, Counterbalanced, and Non-Equivalent Groups Anchor Test (NEST) Designs. KE allows a unified discussion of both Chain Equating and Post-Stratification Equating (frequency estimation and Tucker equating) for the NEAT Design. The registration fee for this session includes a copy of the book *The Kernel Method of Test Equating* (2003), by Alina von Davier and Paul Holland.

Session 2: Introduction to the Generalized Graded Unfolding Model and the GGUM2004 Software Program
Presenters: *James S. Roberts, University of Maryland*

Sunday, April 10, 8:30–5:00 Fee: $60

This training session will provide educational research and measurement professionals with an introduction to unfolding item response theory (IRT) models. These models are appropriate when individuals respond to questionnaire items on the basis of proximity rather than dominance. In other words, individuals endorse items that appear close to their own positions or "ideal points." Item responses indicative of a proximity response process have been repeatedly found in attitude, personality and preference measurement. They have also been found when measuring certain individual differences within developmental processes that occur in distinct stages. The session will emphasize models in the family known as "graded unfolded models." This family contains eight alternative IRT models and includes the generalized graded unfolding model (GGUM) of Roberts, Donoghue and Laughlin (2000). The GGUM is the most general of the unidimensional, parametric item response models in the psychometric literature.

The training session will also introduce attendants to the GGUM2004 software. GGUM2004 calculates parameter estimates for all eight models in the GGUM family. It implements a user-friendly Windows interface, provides realtime graphics to assess model performance, and calculates alternative indices of person, item and model fit. The program is distributed freely via the internet**.**

The session will consist of both a didactic presentation and two exemplar computer exercises which will illustrate both the capabilities of the GGUM2004 software, and the instructor's approach to analyzing attitudinal data.

Session 3: Considerations in Setting Performance Standards
Presenters: *Marianne Perie, ETS; Michael Zieky, ETS; Mary Pitoniak, ETS*

Sunday, April 10, 9:00–5:00 Fee: $65

This training session will examine standard setting. It intends to answer questions regarding how to choose a method, which methods are currently being used, and how to know if the cut scores set for an assessment yield valid interpretations within the context of a particular testing program. Beginning with a historical overview, the session will provide a context regarding how decisions about standard setting are made today. Methodologies currently being used by NAEP and the states in setting performance standards will be reviewed. Hands-on practice time will be given to allow participants to thoroughly understand the cognitive tasks involved in making the judgments for two of the most commonly used methods, Bookmark and modified Angoff. This exercise will also prepare participants to plan and run modified Angoff and Bookmark standard setting workshops. Finally, significant time will be devoted to studying the validity of standard-setting procedures and the resulting cut scores. Using Kane's framework, the session will explore three sources of evidence: procedural, internal, and external. This session is intended for anyone who needs to understand how to run a standard-setting session and the complexities involved. Participants will be provided with a booklet containing a series of articles relevant to the field, as well as sample standard setting materials.

Session 4: Latent Variable Models for Cognitive Diagnosis
Presenters: *Jeffrey Douglas, University of Illinois; Hua-Hua Chang, University of Texas; Louis Roussos, University of Illinois; Andre Rupp, University of Ottawa; Jimmy de la Torre, Rutgers University; Curtis Tatsuoka, George Washington University; Kimuki Tatsuoka; Teachers College, Columbia University*

Sunday, April 10, 9:00–4:15 Fee: $95

The development of Rule Space as well as several multidimensional latent class models for cognitive diagnosis presents the opportunity to develop and analyze tests in ways that reveal information with more diagnostic value, when compared with traditional approaches. In these methods for cognitive diagnosis, mastery of particular skills or states of knowledge can be represented by a list of binary latent variables, indicating mastery of each of a finite set of skills under diagnosis. The main objective of cognitive diagnosis is to classify examinees according to this list of skills. In this training session, several popular modeling and classification approaches will be discussed. Rule Space represents an early and lasting approach to this problem, and will be discussed in depth. Three conjunctive latent class models known as the DINA, NIDA, and Fusion models will then be considered, along with issues of model estimation and goodness-of-fit. Because of the multidimensional nature of these models, estimation benefits greatly if it can adapt to previous responses. To address this, computerized adaptive testing (CAT) is considered. Because Fisher information does not apply to discrete latent variables, alternative and computationally simple item selection rules are introduced. For CAT settings in which both traditional and diagnostic models are being used, CAT algorithms are introduced for ensuring reliable information for these dual objectives. The training session concludes with a discussion of practical implementation issues and the inferential limitations of the models.

Session 5: Accommodations—Measurement and Policy Perspectives: Elementary Through Postsecondary
Presenters: *Noel Gregg, University of Georgia; Martha Thurlow, University of Minnesota; Nicole Ofiesh, University of Arizona; Cara Cahalan, ETS; Linda Cook, ETS*

Sunday, April 10, 1:00–5:00 Fee: $35

The purpose of this session will be to discuss state-of-the art research on-going in the area of accommodations (e.g., different item functioning [DIF], portfolio assessment, extended time, and alternative text), national response to the accommodation of diverse learners, policy challenges, and critical measurement issues.

Session 6: Principles and Applications of Interrater Reliability
Presenters: *Jamal Abedi, CRESST/UCLA*

Sunday, April 10, 1:00–5:00 Fee: $20

Different approaches for estimating interrater (scorer) reliability will be introduced, their limitations will be discussed, and their application will be illustrated. Additionally, generalizability of raters will be discussed using a typical two-facet rater ×task design. A copy of software (ITRS) performing interrater reliability and a two-facet design generalizability developed by the author will be distributed to the audience.

Session 7: ABCs of Automated Essay Scoring
Presenters: *Mark D. Shermis, Florida International University; Scott Elliot, Vantage Learning, Inc.*

Sunday, April 10, 1:00–5:00 Fee: $85

This session is designed to take participants in a step-by-step approach to creating and implementing a statistical scoring model for automated essay scoring (AES). The process will be illustrated using the Vantage learning product Intellimetric™ and MyACCESS. Following a brief overview of the history, theory of AES, and the four major AES systems available, the process will begin by designing a prompt, collecting pilot information, having human raters score essays at different cut points, creating a model, validating it, and using it operationally for either scoring or instructional purposes. Aspects of Intellimetric™ and MyACCESS will be demonstrated to show how AES can be implemented for writing interventions and the types of writing interventions currently underway that use AES as an instructional tool. Information from other AES developers will be distributed as well.

Session 8: Multidimensional Item Response Theory
Presenters: *Brian Habing, University of South Carolina; Amy G. Froelich, Iowa State University*

Sunday, April 10, 9:00–4:30 Fee: $90

One common problem in educational measurement is determining if an exam or scale satisfies the twin assumptions of unidimensionality and local independence. When these assumptions fail it is then necessary to examine the underlying multidimensional/locally dependent structure and either model that structure or refine the original scale. This training session is designed for those who have been exposed to the standard 1PL and 3PL IRT models and deal with (potentially) multidimensional educational assessments or surveys. It focuses on developing an intuitive understanding of the concepts and method as opposed to rigorously developing the mathematics. The session begins with a brief review of the assumptions of local independence and unidimensionality, an overview of multidimensional IRT models (including NOHARM and testlet models), and a survey of the common procedure for testing unidimensionality. Mokken scaling and the conditional covariance methods (DIMTEST, DETECT, HCA-CCPROX) are then examined in detail, with hands-on opportunities to try the procedure on real data sets. Participants will be provided copies of the software used and are encouraged to bring a laptop running Windows 95 or better. Sijtsma and Molenaar's (2002) *Introduction to Nonparametric Item Response Theory* will be provided to the participants along with copies of the key articles referenced.

Session 9: Generalizability Theory and Applications
Presenters: *Robert L. Brennan, University of Iowa; Xiahong Gao, ACT, Inc.*

Monday, April 11, 9:00–5:00 Fee: $90

Generalizability theory liberalizes and extends classical test theory. In particular, generalizability theory enables an investigator to disentangle multiple sources of error measurements. Consequently, generalizability theory is applicable to a broad range of measurement, evaluation, and testing studies that arise in education. 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 consideration of writing assessments and other types of performance assessments. Prerequisites include knowledge equivalent to one course in educational measurement and familiarity with ANOVA at the level treated in introductory graduate courses in education and psychology. A book written by the director and entitled *Generalizability Theory* will be distributed to participants and used as a principle reference in the training session. Computer programs for performing generalizability analyses will be discussed and illustrated.

Session 10: Fundamentals of Polytomous Response IRT Models and Applications
Presenters: *Ronald K. Hambleton, University of Massachusetts at Amherst; Lisa A. Keller, University of Massachusetts at Amherst*

Monday, April 11, 9:00–4:00 Fee: $85

Recently, we have seen substantially more use of polytomously scored items in educational and psychological assessment. Performance tasks, writing samples, surveys to assess attitudes, quality-of-life measures, etc. are all associated with polytomous scoring of the candidate response data. Unfortunately, while many of the popular polytomous IRT models being used today, and the software too, have been available for some time, only a small number of measurement specialists know how to use these models. Those without the knowledge are not even able to understand the psychometric work being done, or provide valid criticism when appropriate. Instructional materials are not readily available either to help practitioners, as they are for analyzing binary-scored data with IRT models. This workshop is being planned as a model first step in changing the current state of knowledge and skills among many graduate students, instructors, and testing specialists.

We propose (1) to describe many of the popular unidimensional polytomous response IRT models (e.g., partial credit, generalized partial credit, and graded response models), their assumptions, and address model parameters interpretations; (2) to explain how parameter estimation and model fit can be carried out; (3) to describe available software and demonstrate the use of that IRT software (e.g., Parscale), and (4) to describe applications to test development and equating via specific example using real data.

Session 11: Grading
Presenters: *Susan M. Brookhart, Duquesne University; James H. McMillan, Virginia Commonwealth University*

Monday, April 11, 9:00–12:00 Fee: $45

Participants in this training session will learn how concepts in educational measurement apply in the context of grading. Sometime, NCME members who have been trained in mainly large-scale measurement principles and procedures find themselves with a smaller repertoire of sound explanations and examples when they are called upon to deal with classroom measurement. The goal of this session is to enlarge that repertoire in the area of grading. Interest in grading is increasing in the current climate of educational accountability.

Session 12: Examining Validity Issues in State Accountability Systems
Presenters: *Brian Gong, National Center for the Improvement of Educational Assessment; Ellen Forte Fast, edCount; and Art Coleman, Holland & Knight*

Monday, April 11, 8:00–12:00 Fee: $65

Performance-based educational accountability systems are meant to achieve highly valued educational goals, such as universal literacy. They are designed to function by imposing consequences for performance, including interventions and sanctions for performance below a specified target and, in some cases, rewards for performance well above a target. These consequences are meant to elicit changes in how a school or district functions; for example, the threat of sanction may motivate a district to revisit and improve the alignment between its curricula and the state standards or offer enhanced professional development programs to improve the quality of instruction. In spite of their apparent logic and well intended purposes, however, it cannot be assumed that accountability systems actually work or are doing more good than harm—even if test scores rise after their implementation. Rather, the quality of the data on which decisions are based, the quality of the decisions, the effectiveness of the sanctions, interventions, and rewards, and the avoidance of negative unintended consequences must all be examined as part of an organized process for supporting the validity of the system. This session will clarify the reasons why this work is critical and provide a framework and examples to guide participants in the design and implementation of their own validation processes. Experience with accountability systems is the only prerequisite. The registration fee for this session includes a copy of the book *A Framework for Examining Validity in State Accountability Systems,* recently published by the Council of Chief State School Officers.

Session 13: Preparing Tests With Test Security in Mind
Presenters: *James C. Impara, Caveon; David Foster, Caveon*

Monday, April 11, 8:00–12:00 Fee: $45

States and local school districts have increased the frequency with which they are involved in high-stakes testing. More and more educational entities have been affected by NCLB and have instituted their own high-stakes testing programs related to end-of-course exams, and high school graduation. Similarly, high-stakes credentialing (both licensure and certification) testing is also on the increase. As the stakes association with testing increase, so do the rewards for test fraud. Test fraud takes on many forms: cheating, and stealing test items to use in test prep courses or to sell on the Internet. What can we, as test developers or sponsors, do to help thwart those individuals who may be involved in test fraud?

This course provides insights and strategies to help those in the testing industry develop tests that hinder those who would behave in fraudulent ways. Test fraud is defined, strategies for writing both multiple-choice and performance items that will make fraudulent behavior more difficult are demonstrated, and test assembly strategies to help protect against fraud are discussed.

Eight multiple-choice item-writing strategies and the type of test fraud these strategies help to eliminate are discussed. Six performance- related item development strategies are divided into two categories: general strategies and strategies for computer-based testing. Four test development and/or test assembly methods are also included in the discussion of how to make tests more secure.

Session 14: A Nonlinear Mixed Models Approach to IRT
Presenters: *Paul De Boeck, K. U. Leuven, Belgium; Mark Wilson, UC Berkeley; Frank Rijmen, K. U. Leuven, Belgium; Francis Tuerlinckx, K. U. Leuven, Belgium*

Monday, April 11, 9:00–2:45 Fee: $65

The central message of the introduction is that it is beneficial to see IRT models as extensions of linear regression models in the case where the data are repeated categorical observations. While the link function and the random component of the regression model remain the same, the most interesting part of the extension concerns the structural part of the model: (1) the kind of predictive function (linear or nonlinear, e.g., bilinear), and (2) the effects (weights) of the predictors (fixed effects or mixed effects—the combination of fixed and random). Then we will discuss and illustrate how multilevel modeling and structural equation modeling (SEM) for categorical data can be expressed from the perspective of nonlinear mixed modeling and vice versa. This will be illustrated with various software-related approaches for multilevel analysis and SEM. Then, starting from some well known models, other and less well known models will be framed in this approach, based on a volume recently published by Springer: *Explanatory Item Response Models: A Generalized Linear and Nonlinear Mixed Approach* (De Boeck & Wilson, 2004). We will illustrate how the models can be estimated with the SAS procedure NLMIXED and WinBUGS. Last, some new models will be presented to illustrate the flexibility of the approach (e.g., SEM with latent person variables and latent item variables, models with gradual change from a random turning point, and copula models for local item dependency).

Session 15: Tips for Graduate Students: Advice for Finishing School, Obtaining a Job, and Starting a Career
Presenters: *Deborah J. Harris, ACT, Inc.*

Monday, April 11, 8:00–12:00 Fee: $7.50

The training session has three main components:
1. Finishing up the Ph.D., including finding a dissertation topic and how to maximize experiences while still a student (classes, internships, work experiences, networking, professional associations);
2. Obtaining a job, including how to locate where jobs are available (universities, testing companies, school districts, state departments, professional/licensing organizations, etc.), how to apply for jobs (including targeting cover letters, references, and resumes) and the interview process; and
3. Beginning a career, including job politics, adjusting to the environment, career path, publishing, professional service, being a mentor/finding a mentor, balancing work and life, and what if I hate my job.

Session 16: Teaching Educational Statistics: A Guided Discussion of 10 Propositions
Presenters: *Ronald C. Serlin, University of Wisconsin; Michael A. Seaman (moderator), University of South Carolina; Patricia Busk, University of San Francisco; Sharon Weinberg, New York University; and Gabriella Belli, Virginia Tech*

Monday, April 11, 2:15–5:15 Fee: $20

In this session, five discussants, including one discussant/moderator, will interact with one another and with members of the audience in a discussion of 10 propositions that could fundamentally govern instructional decisions in the preparation and teaching of educational statistics courses. This session will not be a how-to or a methods sharing session, but will instead focus on basic principles for teaching educational statistics. These principles are derived from theory, research, and the experiences of the discussants. For each proposition, a counterpoint will be presented to promote thought, potential debate, and a lively discussion. Session attendees will receive source information for the referenced theories and research.

Session 17: The History of Educational Testing in the United States
Presenters: *H. D. Hoover, University of Iowa*

Monday, April 11, 2:15–6:15 Fee: $7.50

This session will focus on the history of achievement and ability testing in the United States. The primary emphasis will be on the evolution of large-scale achievement testing during the 20th century and its impact on American schools. Included will be the early development of standardized achievement tests and the reasons for their widespread use, such as the development of the optical scanner, Title I, the minimum competency testing movement, and the use of tests for political purposes. The ways in which adaptations of these tests like the National Merit Scholarship Qualifying Test and the GED came into widespread use will also be discussed. The establishment of the National Assessment of Educational Progress (NAEP) and the subsequent political demands placed on the “nation’s yardstick” will be addressed.

The use of ability tests in schools will also be included, along with a history of college admissions testing exemplified by the *SAT* and *ACT.* The session will conclude with a discussion of possible changes in educational testing in the 21st century.

Session 18: Planning and Developing a Testing Program
Presenters: *Stephen Downing, University of Illinois at Chicago; Thomas Haladyna, Arizona State University*

Monday, April 11, 2:15–6:15 Fee: $40

This 4-hour training helps interested participants learn to plan and develop a high-quality testing program that measures educational achievement in the schools or for credentialing. The training draws heavily from the *Handbook of Test Development* (Downing & Haladyna, in press) which contains 33 chapters on various aspects of test development authored by leading authorities in testing. A 12-step process forms the basis for the planning of a testing program. Participants are expected to put into practice concepts, principles, and procedures learned from this training session to create a hypothetical or real test plan. This document can be used to develop a testing program that follows guidelines suggested in the *Standards for Educational and Psychological Testing.* The plan should also enable the collection and organization of validity evidence that is valuable in any testing program.