



*2009 GMAC® Conference on Computerized Adaptive Testing*

**Radisson University Hotel–Minneapolis• Minneapolis, Minnesota**

*Tuesday, June 2nd, and Wednesday, June 3rd, 2009*



On behalf of the many individuals at GMAC® and the University of Minnesota who have worked hard to organize this meeting, we welcome you to the 2009 GMAC Conference on Computerized Adaptive Testing. Computerized adaptive testing is one of the more exciting aspects of educational and psychological measurement. Here practitioners and theorists get to apply advanced mathematical models to deliver state-of-the-art assessments that are responsive to the needs of both the test developer and the test taker.

As you look through the program, we are sure you will be struck by the range of topics covered by the presentations and posters. GMAC and the University of Minnesota have been able to attract many leaders in the field of measurement. As conference attendees, we will have an opportunity to reflect on actual computerized adaptive testing practices, learn more about leading programs, benefit from the hard learned lessons of our colleagues, and ponder the possibilities from leading edge research in the field. As you can see, this is a very applied conference.

We are certain you will also be struck by the international representation at this meeting. We extend a special welcome to our colleagues who have traveled from Brazil, Canada, India, Israel, Japan, Korea, Russia, Singapore, Spain, Taiwan, The Netherlands, and other locations around the world. Clearly, CAT is an important and fascinating topic globally, and we are fortunate that our international colleagues are able to share their work with us.

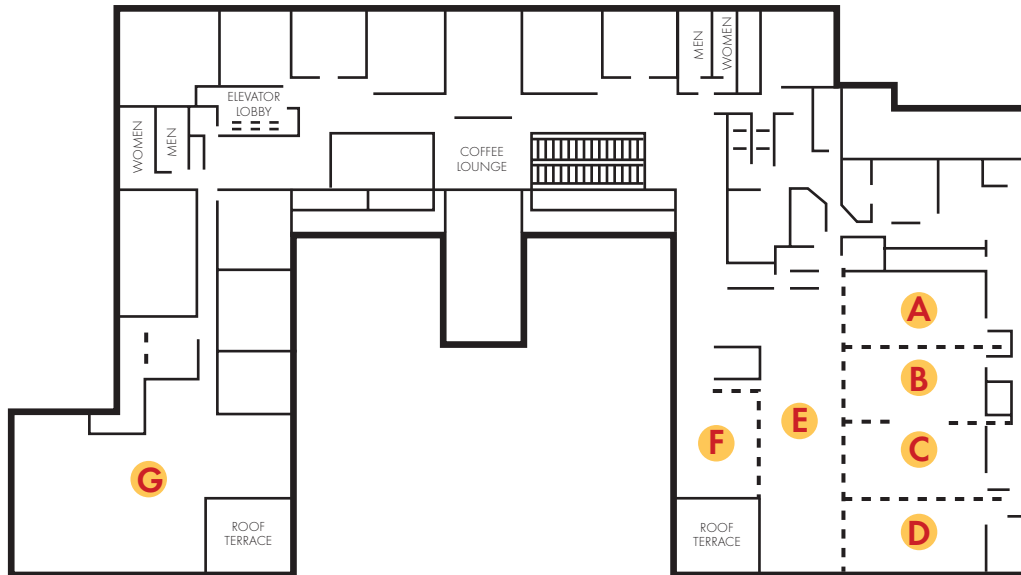
A special feature of this conference is the opportunity for formal and informal discussions with each other after papers are presented, between sessions, and at the reception tonight. The goal is to have rich, and possibly heated, discussions on key topics. We intentionally choose to leave time to the audiences to discuss papers and not to have assigned discussants. Please be ready to share your experiences, discuss your own research, and present your own views. It is the program committee's hope that these discussions will be among the best parts of the meeting.

Again, welcome to the 2009 GMAC Conference on Computerized Adaptive Testing. We sincerely hope this will be a stimulating, productive, and rewarding meeting for us all.

Lawrence M. Rudner, Graduate Management Admission Council®  
David Weiss, University of Minnesota at Twin Cities

## Radisson University Hotel, Second Floor

Radisson University Hotel–Minneapolis • 615 Washington Avenue S.E. • Minneapolis, Minnesota



- A** University Ballroom Section A
- B** University Ballroom Section B
- C** University Ballroom Section C
- D** University Ballroom Section D
- E** Prefunction Area
- F** Faculty Room
- G** Humphrey Ballroom

## MONDAY, JUNE 1

### PRE-CONFERENCE SESSIONS (Separate registration required.)

8:00 am–5:00 pm	Registration
<i>Prefunction Area</i>	
9:00–10:30 am	<b>IRT for CAT</b> (Including: basic assumptions of CAT, IRT models and their parameters, $\theta$ estimation and standard error, estimating item parameters, information for dichotomous and polytomous models, and linking item parameters to create an item bank)
<i>University Ballroom, Sections C &amp; D</i>	
10:30–10:45 am	Refreshment Break
<i>Prefunction Area</i>	
10:45 am–12:15 pm	<b>IRT for CAT (cont.)</b> (Including: basic assumptions of CAT, IRT models and their parameters, $\theta$ estimation and standard error, estimating item parameters, information for dichotomous and polytomous models, and linking item parameters to create an item bank)
<i>University Ballroom, Sections C &amp; D</i>	
12:15–1:45 pm	Lunch
<i>University Ballroom, Sections A &amp; B</i>	
1:45–3:15 pm	<b>Essentials of CAT</b> (Including: basic elements of a CAT, types of CAT, CAT versus sequential testing, what to consider before implementing CAT, implementing a live CAT, and operational issues for some CATs)
<i>University Ballroom, Sections C &amp; D</i>	
3:15–3:30 pm	Refreshment Break
<i>Prefunction Area</i>	
3:30–5:00 pm	<b>Essentials of CAT (cont.)</b> (Including: basic elements of a CAT, types of CAT, CAT versus sequential testing, what to consider before implementing CAT, implementing a live CAT, and operational issues for some CATs)
<i>University Ballroom, Sections C &amp; D</i>	

## TUESDAY, JUNE 2

7:00–8:30 am	Breakfast Buffet
<i>University Ballroom, Sections A &amp; B</i>	
8:00 am–5:00 pm	Registration
<i>Prefunction Area</i>	
8:30–9:00 am	<b>Conference Opener</b> Lawrence M. Rudner, Graduate Management Admission Council® David J. Weiss, University of Minnesota at Twin Cities
<b>WELCOME</b> <i>University Ballroom, Sections C &amp; D</i>	

# Agenda-at-a-Glance

## TUESDAY, JUNE 2

9:00–10:15 am

*University Ballroom,  
Sections C & D*

### Realities of CAT

**Chair:** David J. Weiss, University of Minnesota at Twin Cities

#### *Effect of Early Misfit in Computerized Adaptive Testing on the Recovery of $\theta$*

Rick Guyer and David J. Weiss, University of Minnesota

#### *Quantifying the Impact of Compromised Items in CAT*

Fanmin Guo, Graduate Management Admission Council

#### *Guess What? Score Differences with Rapid Replies versus Omissions on a Computerized Adaptive Test*

Eileen Talento-Miller and Fanmin Guo, Graduate Management Admission Council

#### *Termination Criteria in Computerized Adaptive Tests: Variable-Length CATs Are Not Biased*

Ben Babcock and David J. Weiss, University of Minnesota

10:15–10:30 am

Refreshment Break

*Prefunction Area*

10:30 am–12:00 pm

*University Ballroom,  
Sections C & D*

### CAT for Classification

**Chair:** David J. Weiss, University of Minnesota at Twin Cities

#### *Computerized Classification Testing in More Than Two Categories by Using Stochastic Curtailment*

Theo J.H.M. Eggen and Jasper T. Wouda, CITO, Arnhem, The Netherlands

#### *Utilizing the Generalized Likelihood Ratio as a Termination Criterion*

Nathan A. Thompson, Assessment Systems Corporation

#### *Adaptive Testing Using Decision Theory*

Lawrence M. Rudner, Graduate Management Admission Council

#### *“Black Box” Adaptive Testing by Mutual Information and Multiple Imputations*

Anne Thissen-Roe, Kronos

#### *A Comparison of Computerized Adaptive Testing Approaches: Real-Data Simulations of IRT- and Non-IRT-Based CAT with Personality Measures*

Monica M. Rudick, Wern How Yam, and Leonard Simms, University of Buffalo

12:00–1:00 pm

Lunch

*University Ballroom,  
Sections A & B*

12:30–2:00 pm

**CONCURRENT POSTER  
SESSION**

*Prefunction Area*

### CAT Research and Applications Around the World

#### *A Comparison of Three Methods of Item Selection for Computerized Adaptive Testing*

Denise Reis Costa and Camila Akemi Karino, CESPE/University of Brasilia; Fernando A.S. Moura, Federal University of Rio de Janeiro; Dalton F. Andrade, Federal University of Santa Catarina, Brazil

#### *Adequacy of an Item Pool for Proficiency in English Language from the University of Brasilia for Implementation of a CAT Procedure*

Camila Akemi Karino, Denise Reis Costa, and Jacob Arie Laros, CESPE/University of Brasilia, Brazil

#### *Development of an Item Model Taxonomy for Automatic Item Generation in Computerized Adaptive Testing*

Hollis Lai, Mark J. Gierl, and Cecilia Alves, University of Alberta, Canada

#### *An Approach to Implementing Adaptive Testing Using Item Response Theory in a Paper-Pencil Mode*

V. Natarajan, MeritTrac Services Pvt. Ltd, India

(continued)

## TUESDAY, JUNE 2

12:30–2:00 pm

### CONCURRENT POSTER SESSION

*Prefunction Area*

### CAT Research and Applications Around the World (continued)

#### *Assessing the Equivalence of Internet-Based vs. Paper-and-Pencil Psychometric Tests*

Naomi Gafni, Keren Roded, and Michal Baumer, National Institute for Testing and Evaluation, Israel

#### *Features of a CAT System and Its Application to J-CAT*

Shingo Imai et al., Yamaguchi University, Japan

#### *Adaptive Measurement of Cognitive Ability Based on a Person's Zone of Nearest Development*

Marina Chelyshkova and Victor Zvonnikov, State University of Management, Russia

#### *Implementing Figural Matrix Items in a Computerized Adaptive Testing System: Singapore's Experience*

Tay Poh Hua and Raymond Fong, Ministry of Education, Singapore

#### *Constrained Item Selection Using a Stochastically Curtailed SPRT*

Jasper T. Wouda and Theo J.H.M. Eggen, CITO, The Netherlands

#### *Using Enhanced Effective Response Time to Detect the Extent and Track the Trend of Item Pre-Knowledge on a Large-Scale Computer Adaptive Assessment*

Jie Li and Xiang Bo Wang, ACT, Inc., United States

#### *Computerized Adaptive Testing for the Singapore Employability Skills System (ESS)*

Patrick Rickard, CASAS; James B. Olsen, Alpine Testing Solutions; Debalina Ganguli, CASAS; and Richard Ackermann, Team Code, Inc., United States

#### *Criterion-Related Validity of an Innovative CAT-Based Personality Measure*

Robert J. Schneider, PDRI; Richard A. McLellan, PreVisor, Inc.; Tracy M. Kantrowitz, PreVisor, Inc.; Janis S. Houston, PDRI; Walter C. Borman, PDRI; United States

1:00–1:40 pm

*University Ballroom,  
Sections C & D*

### CAT in Spain and Israel

**Chair:** David J. Weiss, University of Minnesota at Twin Cities

#### *Computerized Adaptive Testing in Spain: Description, Item Parameter Updating and Future Trends of eCAT*

Francisco J. Abad, Universidad Autónoma de Madrid; David Aguado, Universidad Autónoma de Madrid; Juan Ramón Barrada, Universidad Autónoma de Barcelona; Julio Olea, Universidad Autónoma de Madrid; Vicente Ponsoda, Universidad Autónoma de Madrid, Spain

#### *Twenty-Two Years of Applying CAT for Admission to Higher Education in Israel*

Naomi Gafni, National Institute for Testing and Evaluation, Jerusalem, Israel

2:00–3:15 pm

### CONCURRENT SESSION I

*University Ballroom,  
Sections C & D*

### Item Selection

**Chair:** Lawrence M. Rudner, Graduate Management Admission Council

#### *Item Selection and Hypothesis Testing for the Adaptive Measurement of Change*

Matthew Finkelman, Tufts University School of Dental Medicine; David J. Weiss, University of Minnesota; and Gyeonam Kim-Kang, Korea Nazarene University

#### *A Gradual Maximum Information Ratio Approach to Item Selection in Computerized Adaptive Testing*

Kyung (Chris) T. Han, Graduate Management Admission Council

#### *Item Selection with Biased-Coin Up-and-Down Designs*

Yanyan Sheng, Southern Illinois University at Carbondale

#### *A Burdened CAT: Incorporating Response Burden with Maximum Fisher's Information for Item Selection*

Richard J. Swartz, The University of Texas M.D. Anderson Cancer Center, and Seung W. Choi, Northshore University Health System Research Institute and Northwestern University

# Agenda-at-a-Glance

## TUESDAY, JUNE 2

2:00–3:15 pm

### CONCURRENT SESSION II

*Faculty Room*

#### Real-Time Analysis

**Chair:** Fanmin Guo, Graduate Management Admission Council

*Adaptive Item Calibration: A Simple Process for Estimating Item Parameters Within a Computerized Adaptive Test*

G. Gage Kingsbury, Northwest Evaluation Association

*On the Fly Item Calibration in Low Stakes CAT Procedures*

Sharon Klinkenberg, Department of Psychology, University of Amsterdam; Marthe Straatemeier, Department of Psychology, University of Amsterdam; Gunter Maris, CITO; and Han van der Maas, Department of Psychology, University of Amsterdam

*An Automatic Online Calibration Design in Adaptive Testing*

Guido Makransky, University of Twente/Master Management International A/S, and Cees A. W. Glas, University of Twente

*Investigating Cheating Effects on the Conditional Simpson and Hetter Online Procedure with Freeze Control for Testlet-Based Items*

Ya-Hui Su, University of California, Berkeley

3:15–3:25 pm

Refreshment Break

*Prefunction Area*

3:25–5:30 pm

*University Ballroom,  
Sections C & D*

#### US GOVERNMENT-SUPPORTED CAT PROGRAMS AND PROJECTS

**Chair:** David J. Weiss, University of Minnesota at Twin Cities

##### Department of Defense

*The Nine Lives of CAT-ASVAB: Innovations and Revelations*

Mary Pommerich, Daniel O. Segall, and Kathleen E. Moreno, Defense Manpower Data Center

##### National Institutes of Health

*The CAT-DI Project: Development of a Comprehensive CAT-Based Instrument for Measuring Depression*

Robert D. Gibbons, University of Illinois at Chicago

*Development of a CAT to Measure Dimensions of Personality Disorder: The CAT-PD Project*

Leonard J. Simms, University of Buffalo

*The MEDPRO Project: An SBIR Project for a Comprehensive IRT and CAT Software System*

*IRT Software* – David Thissen, The University of North Carolina at Chapel Hill and Scientific Software International

*CAT Software* – Nathan Thompson, Assessment Systems Corporation

7:00–9:00 pm

Reception

*Humphrey Ballroom*

## WEDNESDAY, JUNE 3

6:00–8:00 am

Breakfast Buffet

*University Ballroom,  
Sections A & B*

8:15–9:25 am

### CONCURRENT SESSION III

*University Ballroom,  
Sections C & D*

#### Item Exposure

**Chair:** Lawrence M. Rudner, Graduate Management Admission Council

*Reviewing Test Overlap Rate and Item Exposure Rate as Indicators of Test Security in CATs*

Juan Ramón Barrada, Universidad Autónoma de Barcelona; Julio Olea, Vicente Ponsoda, and Francisco J. Abad, Universidad Autónoma de Madrid, Spain

*Optimizing Item Exposure Control and Test Termination Algorithm Pairings for Polytomous Computerized Adaptive Tests with Restricted Item Banks*

Michael Chajewski and Charles Lewis, Fordham University

*Limiting Item Exposure for Key-Difficulty Ranges in a High-Stakes CAT*

Xin Li, Kirk A. Becker, and Jerry L. Gorham, Pearson VUE; Ada Woo, National Council of State Boards of Nursing

8:15–9:25 am

### CONCURRENT SESSION IV

*Faculty Room*

#### Multidimensional CAT

**Chair:** Nate Thompson, Assessment Systems Corporation

*Comparison of Adaptive Bayesian Estimation and Weighted Bayesian Estimation in Multidimensional Computerized Adaptive Testing*

Po-Hsi Chen, Taiwan Normal University

*Comparison of Ability Estimation and Item Selection Methods in Multidimensional Computerized Adaptive Testing*

Qi Diao and Mark Reckase, Michigan State University

*Multidimensional Adaptive Testing: The Application of Kullback-Leibler Information*

Chun Wang and Hua-Hua Chang, University of Illinois at Urbana-Champaign

*Multidimensional Adaptive Personality Assessment: A Real-Data Confirmation*

Alan D. Mead, Avi Fleischer, and Jessica D. Sergent, Illinois Institute of Technology

9:35–10:45 am

*University Ballroom,  
Sections C & D*

#### Item and Pool Development

**Chair:** Lawrence M. Rudner, Graduate Management Admission Council

*A Comparison of Three Procedures for Computing Information Functions for Bayesian Scores from Computerized Adaptive Tests*

Kyoko Ito, Human Resources Research Organization; Mary Pommerich, Defense Manpower Data Center; and Daniel O. Segall, Defense Manpower Data Center

*Adaptive Computer-Based Tasks Under an Assessment Engineering Paradigm*

Richard M. Luecht, The University of North Carolina at Greensboro

*Security and Memorability of Innovative Items*

Anne Wendt, National Council of State Boards of Nursing; Shu-chuan Kao, Pearson VUE; Jerry Gorham, Pearson VUE; and Ada Woo, National Council of State Boards of Nursing

*Evaluation of a Hybrid Simulation Procedure for the Development of Computerized Adaptive Tests*

Steven W. Nydick and David J. Weiss, University of Minnesota

10:45–11:00 am

Refreshment Break

*Prefunction Area*

---

# Agenda-at-a-Glance

---

## WEDNESDAY, JUNE 3

11:00–11:55 am

*University Ballroom,  
Sections C & D*

### **Diagnostic Testing**

**Chair:** Lawrence M. Rudner, Graduate Management Admission Council

#### ***Computerized Adaptive Testing for Cognitive Diagnosis***

Ying Cheng, University of Notre Dame

#### ***Obtaining Reliable Diagnostic Information through Constrained CAT***

Hua-Hua Chang, Jeff Douglas, and Chun Wang, University of Illinois

#### ***Applying the DINA Model to GMAT Focus Data***

Alan Huebner, Xiang Bo Wang, and Sung Lee, ACT, Inc.

---

11:55 am–12:30 pm

*University Ballroom,  
Sections C & D*

### **Wrap-Up and Future Directions**

Lawrence M. Rudner, Graduate Management Admission Council

David J. Weiss, University of Minnesota at Twin Cities

---

## Effect of Early Misfit in Computerized Adaptive Testing on the Recovery of $\theta$

Rick Guyer and David J. Weiss, University of Minnesota

This study focused on how early person misfit affected the recovery of  $\theta$  for a computerized adaptive test (CAT) based on the three-parameter logistic model. Number of misfitting items, generating  $\theta$ , item selection method, and estimation method were independent variables in this study. The number of misfitting initial item responses was varied from  $k = 0$  to 4 items. Ten different generating  $\theta$  values were used at intervals from  $-3$  to  $+3$ . For the five conditions in which  $\theta$  was less than or equal to 0, the first  $k$  responses were fixed to be correct; for the five conditions where  $\theta$  was greater than or equal to 0, misfit was introduced by fixing the first  $k$  item responses to be incorrect. Maximum likelihood, weighted likelihood (WLE), and *expected a posteriori* (EAP) estimation were used to estimate  $\theta$ . Both Fisher information and Kullback-Leibler information item selection methods were used. All independent variables were crossed in the simulation design, with 1,000 simulees per cell. Recovery of  $\theta$  was indexed by bias, standard error, and root-mean-square error at CAT lengths of 15, 25, 35, and 50 items. ANOVA was used to analyze the results and major effects were identified by eta-squared.

It was found that CAT could recover from misfit-as-correct-responses (MCR) for low ability simulees given a sufficient number of items. CAT could not recover from misfit-as-incorrect-responses (MIR) for high ability simulees, even after 50 items. At 50 items, a small amount of bias was observed for 1 misfitting item; as the number of misfitting items increased to 4, the bias increased and was substantial for all positive values of  $\theta$ . The differences between the Fisher and Kullback-Leibler information-based item selection dissipated after 15 items were administered, with one exception: for the MIR conditions, it was found that WLE functioned differently under the two item selection methods even after 50 items were administered. A follow-up study was performed, and it was found that WLE was highly sensitive to item difficulty early in the CAT. Implications of the results and suggestions for future research will be provided.

**For further information:** [guyerr@assess.com](mailto:guyerr@assess.com) or [guyer005@umn.edu](mailto:guyer005@umn.edu)

## Quantifying the Impact of Compromised Items in CAT

Fanmin Guo, Graduate Management Admission Council

If a few test items should become compromised, their impact on test scores would not be constant across different computerized adaptive test (CAT) programs. For the same number of items compromised, the impact might be more serious in some CAT programs than others because the impact interacts with the complexity of the test specification, size of CAT pools, item exposure control, item selection algorithm, and scoring method employed in CAT programs. As a result, evaluating the impact of compromised items on test scores in a CAT program is not easy. Most of the previous research focused on the impact on a group of examinees through simulations.

In this study, a new method of simulation is introduced that focuses on the impact on individual examinees using the GMAT® CAT as an example. For each simulee, two paths of simulations were run. The first path is the conventional simulation under no compromised item condition. The second path follows the selected items and response patterns in the first path until a “compromised” item is “administered.” Then the answer to this item is reset to a correct answer to simulate a “security breach.” After that, the path branches to selecting new items. All the answers to subsequent “compromised” items are set as correct answers until the end of the test. The purpose of this method is to quantify the impact of compromised items as well as its interaction with the item selection and other CAT operational configurations. Since each simulee will have two scores from the two separate paths, this method allows estimating the range of score gains and the number of compromised items seen by each individual. It allows reports that show if  $n$  items from a CAT pool were exposed to  $m$  examinees,

# Conference Abstracts

---

$x$  examinees would gain  $y$  score points due to the impact of compromised items. The method employed in this study applies to any CAT program.

**For further information:** [fguo@gmac.com](mailto:fguo@gmac.com)

## Guess What? Score Differences with Rapid Replies versus Omissions on a Computerized Adaptive Test

*Eileen Talento-Miller and Fanmin Guo, Graduate Management Admission Council*

Estimation of ability in computerized adaptive testing (CAT) relies on the assumption that examinees are responding based on their content knowledge and skills. Guessing might have differential consequences on scoring depending on the situation. In the case of time constraints, examinees are faced with a choice of leaving questions blank or randomly responding. The current study provides guidance for examinees based on real data from an operational CAT. Previous research provides an incomplete picture of the effects of choosing a guessing strategy versus omitting items in the scoring of an operational CAT. The study expands on previous research by using operational as opposed to simulated data, comparing results in verbal and quantitative sections of a test, and framing the results to provide guidance for examinees.

In this study, scores from tests with responses classified as random guesses are compared to scores that would be observed if the items had not been reached. Items are classified as guesses by examining the distribution of latency for correct responses to determine a rapid guessing threshold. The threshold is then checked against the proportion of correct responses at that level to find close to chance levels of correctly answering the item. The guessing threshold of 10 seconds for verbal items and 7 seconds for quantitative items was applied to all item positions. Only consecutive rapid guesses at the end of the section were examined. Scores of examinees who guessed are recalculated to reflect ending the test and omitting the remaining items rather than guessing. Although the results tend to favor guessing as a strategy, the degree of difference varied based on section content, number of items involved, and estimated ability of the examinee. In the verbal section of the test, few differences existed between guessing scores and omission scores. In the quantitative section, the benefit of guessing became more pronounced as the number of items increased. The results are particularly intriguing when ability groups are compared. Both the verbal and quantitative sections show a slight preference for the omission strategy in the low ability group. For the high ability group, apparently severe penalties for omissions in the shorter quantitative measure appear to make guessing the unequivocal strategy of choice. Future research could include more definitive methods for determining random guessing and examinee guessing at different positions within the test rather than merely at the end. Ultimately, the advice for candidates remains the same for a CAT as it would for other tests: Time management is important to allow ample opportunity to give thought to every question.

**For further information:** [talento-miller@gmac.com](mailto:talento-miller@gmac.com)

## Termination Criteria in Computerized Adaptive Tests: Variable-Length CATs Are Not Biased

*Ben Babcock and David J. Weiss, University of Minnesota*

This simulation study examined the performance of several CAT termination rules: four basic termination rules (standard error, minimum information, change in  $\theta$ , and fixed length) and two combinations of standard error and minimum information termination. Four item banks were used: a flat information bank with 500 items, a peaked information bank with 500 items, a flat information bank with 100 items, and a peaked information bank with 100 items. Maximum likelihood scoring was used to estimate  $\theta$ . For non-mixed response vectors,

was incremented by 0.5. In addition to examining the performance of these termination criteria, the study was concerned with further examining the conclusion from previous research that variable-length CATs are more biased than fixed-length CATs (Chang & Ansley, 2003; Yi, Wang, & Ban, 2001). First, a number of variable-length CAT conditions were simulated. Then, the mean number of items administered for selected variable-length conditions was determined and fixed-length CATs were simulated with the appropriate number of items in order to properly compare variable- and fixed-length CATs. CAT performance was compared in terms of test length, as well as bias, RMSE, and correlation in the recovery of true  $\theta$ .

As expected, longer CATs yielded more accurate estimation no matter which termination criterion was used, but there were diminishing returns with a large number of items. It is recommended that CATs should administer a minimum number of 15 to 20 items to ensure stable measurement. The standard error termination rule, also known as the equiprecise measurement rule, performed the best among all the methods if the standard error cutoff was sufficiently low and the item bank contained the amount of Fisher information needed to reach the cutoff. Standard error termination was also quite efficient by administering relatively few items. Change in  $\theta$ , a newer termination criterion, performed slightly worse than its fixed-length termination counterpart. Hybrid termination rules, such as combining minimum information and standard error termination, functioned best when the item bank was small but had a peaked information function. The fixed-length CATs did not perform better than their standard error termination counterpart when equated for average test length. Previous findings stating that variable-length CATs are more biased than fixed-length CATs were the result of two procedural artifacts in prior research: (1) variable-length CATs were generally much shorter than the fixed-length CATs; and (2) most previous studies used Bayesian scoring which biased the shorter variable-length CATs in the previous studies because the prior has a greater effect on estimation when there is less psychometric information. Standard error termination actually performed slightly better than fixed-length CATs of comparable mean length in estimating low true  $\theta$  values.

**For further information: [babco062@umn.edu](mailto:babco062@umn.edu)**

## Computerized Classification Testing in More Than Two Categories by Using Stochastic Curtailment

*Theo J. H. M. Eggen, CITO and University of Twente, The Netherlands*

*Jasper T. Wouda, CITO, The Netherlands*

When classification into a limited number of categories is the main purpose of testing, algorithms based on the application of sequential statistical testing have shown to be better performing alternatives above traditional estimation based computerized adaptive tests (e.g. Reckase, 1983; and Eggen & Straetmans, 2000). In these studies, the sequential probability ratio test (SPRT; Wald, 1947) is applied in order to decide whether more observations on items are needed and which classification decision is to be made. When a decision cannot be made with the predetermined decision error rates, in practice the procedure is always truncated at a maximum test length. Recently Finkelman (2003, 2008) proposed an adaptation of stochastic curtailment with which he created an additional stopping rule for the sequential probability ratio test (SPRT). This “stochastically curtailed sequential probability ratio test, or SCSPT, generally follows the same rules as the conventionally truncated SPRT, including its stopping rule. The SCSPT, however, adds some rules in order to be able to stop testing in the cases where a change in decision between categories is possible, but unlikely. Finkelman (2003) introduced the method for the case of classifications in two categories and items selected to be most informative at the classification point. In this paper the generalization of the application of the SCSPT to problems with more than two categories is discussed with a focus on the problems encountered in generalizing to the three-category

# Conference Abstracts

---

problem. In general the (optimal) composition of the test cannot be fixed in advance when there is more than one cutting point, which is a requirement of Finkelman's SCSPT. This report describes the way the application of stochastic curtailment in combinations of SPRTs can be combined with the adaptive item selection in the test. The performance of the proposed procedures is illustrated by results of simulation studies.

**For further information:** [Theo.Eggen@cito.nl](mailto:Theo.Eggen@cito.nl)

## Utilizing the Generalized Likelihood Ratio as a Termination Criterion

*Nathan A. Thompson, Assessment Systems Corporation*

A common application for adaptive testing is to classify examinees into mutually exclusive groups. Currently, the predominant psychometric termination criterion for designing computerized classification tests is the sequential probability ratio test (SPRT; Reckase, 1983) based on item response theory. This operates by formulating a hypothesis test that a given examinee's ability value is equal to a fixed value below ( $\mu_1$ ) or above ( $\mu_2$ ) the classification cut score. Recently, it was demonstrated that the SPRT, which only uses fixed values, is less efficient than a generalized form that tests whether a given examinee's  $\theta$  is below  $\mu_1$  or above  $\mu_2$  (Thompson, 2007). Moreover, this better represents the conceptual purpose of the exam, which is to test whether  $\theta$  is above or below the cut score.

The purpose of this study was to explore the specifications of the new generalized likelihood ratio (GLR). As with the SPRT, the efficiency of the procedure depends on the nominal error rates and the distance between  $\mu_1$  and  $\mu_2$  (Eggen, 1999). Preliminary results suggest that observed error rates are closest to nominally specified error rates when the values of  $\mu_1$  and  $\mu_2$  are approximately 0.1 from the cut score. The study utilized a *Monte Carlo* approach, with 10,000 examinees simulated under each condition. Three levels of nominal accuracy were investigated (90%, 95%, and 99%), as well as 25 values of the difference between the cut score and  $\mu_1$  or  $\mu_2$  (0.00 to 0.50 in increments of 0.2). Additionally, another formulation was investigated that forms the likelihood ratio based on an integration of the likelihood function. This was also suggested by Thompson (2007), but was not accurate due to the asymmetry of the likelihood function when the three-parameter model is used; the left-hand end of the likelihood function is substantially higher than the right-hand end because of the  $c$  parameter. This artificially biases the ratio in the negative direction. Methods of correcting for this are suggested.

**For further information:** [nthompson@assess.com](mailto:nthompson@assess.com)

## Adaptive Testing Using Decision Theory

*Lawrence M. Rudner, Graduate Management Admission Council*

In the introduction to their classic textbook, Cronbach and Gleser (1957) argue that the ultimate purpose for testing is to arrive at classification decisions. Many of today's decisions are indeed binary, e.g., whether to hire someone, whether a person has mastered a particular set of skills, or whether to certify an individual. Categorical, as opposed to continuous, outcomes are also common, e.g., the percentage of students that perform at the basic, proficient, or advanced level in state assessments. Item response theory (IRT) models have been applied to help make classification decisions by laboriously placing individuals on ability scales and then using cut-points to make classifications. IRT models, however, are not always applicable in practical situations. IRT is fairly complex, relies on several fairly restrictive assumptions, requires large calibration samples, and might not make efficient use of test questions when the goal is simple classification. This paper presents an alternative underlying model for adaptive testing using measurement decision theory and then compares those procedures with IRT in terms of classification accuracy using two sets of simulated item response data. The research examines three ways to select items adaptively using decision theory: a traditional decision theory

sequential testing approach (expected minimum cost), information gain (modeled after Kullback-Leibler, 1951), and maximum discrimination. It also examines the use of Wald's (1947) well-known sequential probability ratio test (SPRT) as a test termination rule in this context.

Initial results show that the minimum cost approach was notably better than the best-case possibility for IRT. Information gain, which is based on entropy and comes from information theory, was almost identical to minimum cost. The simple approach using the item that best discriminates between the two most likely classifications also fared better than IRT, but not as well as information gain or minimum cost. Initial results also show that with Wald's SPRT, large percentages of examinees can be accurately classified with very few items. With only 25 sequentially selected items, for example, some 90% of the simulated state-NAEP examinees were classified with 86% accuracy. Clearly, this is a simple yet powerful and widely applicable model. The advantages of this model are many—the model yields accurate mastery state classifications, can incorporate a small item pool, is simple to implement, requires little pre-testing, is applicable to criterion-referenced tests, can be used in diagnostic testing, can be adapted to yield classifications on multiple skills, and should be easy to explain to nonstatisticians. It is the author's hope that this research will capture the imagination of the research and applied measurement communities. The author can envision wider use of the model as the routing mechanism for intelligent tutoring systems. Items could be piloted with a small number of examinees to vastly improve end-of-unit examinations. Certification examinations could be created for specialized occupations with a limited number of practitioners available for item calibration. Short tests could be prepared for teachers to help make tentative placement and advancement decisions. A small collection of items from one test, say state-NAEP, could be embedded in another test, say a state assessment, to yield meaningful cross-regional information.

**For further information: [LRudner@gmac.com](mailto:LRudner@gmac.com)**

## "Black-Box" Adaptive Testing by Mutual Information and Multiple Imputations

*Anne Thissen-Roe, Kronos*

Over the years, most CAT systems have used score estimation procedures from item response theory. IRT models have salutary properties for score estimation, error reporting, and next-item selection. Some testing purposes favor scoring approaches outside IRT, however. Where a criterion metric is readily available and more relevant than the assessed construct, for example in the selection of job applicants, a predictive model might be appropriate (Scarborough & Somers, 2006). Neither IRT scoring nor unidimensional assessment structure can be assumed. Yet, the primary benefit of CAT remains desirable: shorter assessments with minimal loss of accuracy due to unasked items. Without IRT, it remains possible to create a CAT system that produces an estimated score from a subset of available items, recognizes differential item information given the emerging item response pattern, and optimizes the accuracy of the score estimated at every successive item. No information is needed about the internal mechanisms of the scoring algorithm, provided it has certain properties: (1) the score must be discrete or able to be made discrete, such as by application of cut scores or reporting of integer scale scores; (2) the score can be a nominal category; and (3) the degree to which the score changes when a particular item response is given must vary based on the responses to other items. If these conditions are met, the scoring algorithm can be treated as a "black box," with adaptation conducted on the outside. The method of multiple imputations (Rubin, 1987) might be used to simulate plausible scores given plausible response patterns to unasked items (Thissen-Roe, 2005). This method is also capable of rendering an estimate of the error introduced by unasked questions. Mutual information might then be calculated in order to select an optimally informative next item (or set of items). This is related but not identical to the methods of Weissman (2007) for item selection, and Chambless and Scarborough (2001) for feature selection.

# Conference Abstracts

---

Two neural network-centered scoring algorithms serve as structural examples. In early testing, previously observed response patterns to the complete assessments were resampled according to CAT item selection. The reproduced CAT scores were compared to full-length assessment scores. Approximately 95% accurate assignment of examinees to one of three score categories was achieved with a 70% to 80% reduction in median test length. This method of CAT is more computationally demanding than traditional IRT-based approaches, due to the necessity of completely scoring some hundreds or thousands of response patterns per item selected. Factors influencing performance were also examined during early testing. Reducing the number of multiple imputations used is a way to reduce computation time; it appears to impact assignment accuracy less than limiting items presented under a confidence-based stopping rule. Computation time can also be reduced by sacrificing algorithmic simplicity to move repeated computations outside of the “black box;” however, such shortcuts impose a maintenance burden. Mixing “black box” CAT with Internet testing also requires minimizing the data size and frequency of transactions between client and server, for which the simplest algorithm is well suited.

**For further information:** [anne.thissenroe@kronos.com](mailto:anne.thissenroe@kronos.com)

## A Comparison of Computerized Adaptive Testing Approaches: Real-Data Simulations of IRT- and Non-IRT-Based CAT with Personality Measures

*Monica M. Rudick, Wern How Yam, and Leonard Simms, University at Buffalo, State University of New York*

Various approaches have been implemented to create CAT personality assessments. Recent research has focused on item response theory (IRT) for CAT personality measures, although its use is computationally complex and requires certain assumptions to be met that do not always hold for personality measures. As a result, non-IRT-based CAT approaches, such as the countdown method, have also successfully been applied to CAT versions of personality measures. In the countdown method, there is some debate whether classification or full-scores-on-elevated-scales (FSES) methods are more preferable. In addition, it is unclear how order of item administration might impact item savings and the validity of scores. Both IRT- and non-IRT-based methods appear to yield numerous advantages for CAT assessments, most notably time and item savings, and ease of administration. These two methods have yet to be directly compared, however. The purpose of the present study is to compare non-IRT- and IRT-based approaches utilizing real-data CAT simulations on a large diverse sample ( $N = 8690$ ) that completed the Schedule for Nonadaptive and Adaptive Personality (SNAP). The report focuses on the three longest SNAP scales: disinhibition (DIS), negative temperament (NT) and positive temperament (PT). Simulation analyses compared item savings, item and test information, test validity, and fidelity across the IRT- and non-IRT CAT methods. In addition, countdown method simulations examined whether item presentation order impacted the results. Results will have implications for test developers wishing to apply CAT technology to personality measures.

**For further information:** [mmrudick@buffalo.edu](mailto:mmrudick@buffalo.edu)

## A Comparison of Three Methods of Item Selection for Computerized Adaptive Testing

*Denise Reis Costa, Camila Akemi Karino, CESPE/University of Brasilia, Brazil*

*Fernando A. S. Moura, Federal University of Rio de Janeiro, Brazil*

*Dalton F. Andrade, Federal University of Santa Catarina, Brazil*

One of the most important components of CAT is the set of procedures for item selection. Unlike traditional paper-and-pencil tests, adaptive procedures administer items that fit the examinee's level of proficiency. This selection is based both on the characteristics of the items (e.g., item difficulty or discrimination parameters)

and on the estimated proficiency of the examinee. This study is a work-in-progress that aims to evaluate the performance of three different CAT item selection methods. The first one is derived from the maximum information criterion, one of the most popular item selection methods in CAT. The second method is based on the global information method as defined by Chang and Ying (1996), which uses the Kullback-Leibler measure. The third selection method is based on the predictive analysis defined by the expected maximum information criterion proposed by van der Linden (1998). To evaluate the three different methods, the answers of 10 examinees with different skill levels were simulated for an item pool containing 246 items of the Instrumental English test of the University of Brasilia. The resulting database was fit by a three-parameter logistic model on a scale with mean 0.0 and standard deviation of 1.0, later transformed into a mean of 100 and standard deviation of 25. The examinees' iterative proficiencies were estimated using *expected a posteriori* (EAP). An initial analysis of bias and mean square error suggested that all methods performed similarly to estimate examinees' proficiency. Databank-related characteristics, however, might have influenced those measures, since it is not yet an ideal item pool for CAT implementation. With these results, it can be concluded that there is no apparent statistical difference in relation to the proficiency estimation for the three presented methods for the analyzed item bank.

**For further information:** [denise@cespe.unb.br](mailto:denise@cespe.unb.br)

## Adequacy of an Item Pool for Proficiency in English Language from the University of Brasília for Implementation of a CAT Procedure

Camila Akemi Karino, Denise Reis Costa, and Jacob Arie Laros  
CESPE/University of Brasilia, Brazil

The possibility of applying different item sets according to the level of ability of each respondent has stimulated, among other factors, an increasing use of CAT. In spite of the increasing use, this study is one of the first initiatives in this field in Brazil. The item pool used in this study is a database of the proficiency exam in English language that has been used since 2004 by the University of Brasilia. This proficiency exam assesses the student's comprehension of texts in the English language. The exam is a paper-and-pencil test composed of 50 multiple-choice items. The psychometric item quality was verified using classical test theory and item response theory (IRT). The complete item pool consists of 450 items divided into nine test forms. On average, 330 students responded to each test form. The total number of respondents was 2,969. First, each test was analyzed individually and in a second stage the nine tests were calibrated jointly. Of the 450 items, 37 items were common items among test forms. In the individual analyses, 46 items with biserial correlation less than .20 and 80 items with discrimination parameter in the normal IRT metric less than .50 were eliminated. In the joint analysis, another 58 items with an  $\alpha$  parameter less than .50 were eliminated. After the elimination of these items, the joint IRT analysis revealed a mean discrimination parameter of .77 (SD = .20), varying between .49 and 1.67. In relation to the  $b$  parameter, the existence of a substantial variation in difficulty level of the items was observed (varying between -3.56 and 3.23); however, the majority (75%) of the items showed a  $b$  parameter below .10. The median value of parameter  $c$  was .11 (SD = .04) with a range from .03 to .24. After the joint calibration, successive points of the scale were fixed for anchor items and each of these levels was interpreted pedagogically by specialists. The suitability of the item pool for implementation of a CAT procedure was questioned taking into consideration that 44% of the items needed to be eliminated in order to agree with pre-established psychometric criteria. Nonetheless, both the analysis of the item pool and the scale interpretation permit initial studies for the implementation of a CAT procedure. The item pool as well as the scale could be improved by repeated applications of the English exam using a CAT procedure.

**For further information:** [camilaakarino@gmail.com](mailto:camilaakarino@gmail.com)

## Development of an Item Model Taxonomy for Automatic Item Generation in Computerized Adaptive Testing

*Hollis Lai, Mark J. Gierl, and Cecilia Alves, University of Alberta, Canada*

Computerized adaptive testing (CAT) makes tremendous demands on item banks because CATs require large numbers of test items. CATs require these item volumes for three general reasons. First, as test length increases in fixed-length CATs, requirements for test items increase to ensure that test scores are reliable (Wainer & Eignor, 2000). Second, with the emergence of cognitive adaptive tests (e.g., Zhou, Gierl & Cui, 2008), many more skills are measured at a finer grain size. Thus, more test items are required to measure these large numbers of specific skills. Third, item exposure and security concerns demand that item re-use rates be relatively small. That is, CAT requires a large number of unique test items in operational testing situations. One solution that could be developed to address these three issues is to generate many more items. Automatic item generation is an approach to item development where large numbers of offspring items (also called item instances) are generated from a parent item model. Although automatic item generation can potentially create hundreds and even thousands of items, its effectiveness is reliant on the availability of an efficient framework for creating the parent item models. The components in a parent item model for a multiple-choice item consist of the stem (the component of an item that forms the context of the question the examinee is required to answer), the options (a set of alternatives with one correct option and multiple distracters to answer the question), and any auxiliary information (e.g., pictures, graphs).

To identify possible item model types, Gierl, Zhou, and Alves (2008) developed a taxonomy to categorize and delineate the levels of variation in components of the parent item model. One limitation of the study by Gierl et al., however, was that it focused only on mathematics items. To be applied in diverse testing situations, item models need to be created in many different content areas to allow for automatic item generation. The present study will apply the taxonomy to item models from diverse content areas, including Language Arts, Social Studies, and Science, to generate items for a computer-based testing program. While there might have been other implementations of item generation, few have been documented (Irvine, 2002). Hence, the implication of the present study is to demonstrate a systematic way to generate test items that creates large numbers of items in diverse content areas, thereby lowering the cost of item development while maintaining a high level of quality in the development process.

**For further information:** [hollis.lai@ualberta.ca](mailto:hollis.lai@ualberta.ca)

## An Approach to Implementing Adaptive Testing Using Item Response Theory in a Paper-Pencil Mode

*V. Natarajan, MeritTrac Services Pvt. Ltd, India*

In India, as most of the large-scale testing is conducted in the paper-pencil (offline) mode, it is important to arrive at models of implementing item response theory (IRT) in an offline/paper-pencil mode. MeritTrac has experimented with conducting an IRT-based test in a paper-pencil mode for the analytical abilities test for engineering graduates. With the help of item characteristics calculated prior to the test, a six-item test with increasing item difficulty was created as a test form on paper. Research shows that a 6/10 item test normally can be compared to 25 or more items in the test. The test was then administered to the candidates in an offline mode. The responses of the examinees were then entered in student tracking software that had been specially coded for this purpose. The output of this gives an estimation of the examinee's true score as if he/she has taken the parent 25-item test. Since it is not very feasible to conduct an online test everywhere, especially in a country like India, the importance of adaptive testing in offline mode increases manifold. In this model, we

only need a single computer with student tracking software and pre-published test forms consisting of items whose characteristics have been calculated on the basis of past responses. Thus the offline mode is much more practical and is just as accurate as the online mode.

In the analytical abilities test, we have looked at 100 items and the responses of 1,000+ examinees on each of these items, which we entered into BILOG and generated item difficulty values. Ninety-three items were found to be relevant and the parent test of 93 items eventually emerged. The items were grouped into six sets and 10 items were selected (one item each very easy and easy, and two items from below average, average difficulty, and very difficult). Several sets of 10-item adaptive tests were selected and administered to the examinees. Their responses to 10 items were categorized in terms of 9,8,7,6,5,4,3,2,1 correct and a table generated from which ability and true scores can be read. In this methodology, the test administrator needs to be very cautious when dealing with student tracking software so that mistakes are not made in entering the values of item numbers in the reshuffled version and the examinee's responses.

**For further information: [madan@merittrac.com](mailto:madan@merittrac.com)**

## Assessing the Equivalence of Internet-Based vs. Paper-and-Pencil Psychometric Tests

*Naomi Gafni, Keren Roded, and Michal Baumer, National Institute for Testing and Evaluation, Israel*

Few studies have yielded information regarding the equivalence of high-stakes admissions tests administered via the Internet and paper-and-pencil administrations of those tests (Potosky & Bobko, 2004). Despite the lack of evidence regarding the equivalence of scores obtained in these two modalities, there is increasing demand for Internet-based testing, with the number of recruitment and admissions tests administered via the Internet constantly rising. This is largely due to the convenience and efficiency that the medium offers. The Psychometric Test, which is used for admission to institutions of higher education in Israel, is a high-stakes examination. The test consists of three sections: Verbal Reasoning (60 items), Quantitative Reasoning (60 items), and English as a Foreign Language (54 items). All items are in multiple-choice format. At the present time, most of the examinees take the paper-and-pencil version of the test. It is anticipated that Internet-based administration will be expanded. Given that this process will be gradual, and for a period of time the test will be administered in two parallel modalities, establishing the equivalence of scores is of paramount importance.

The goal of the present study was to compare the achievement of examinees who took the paper-and-pencil version of the Psychometric Test with the achievement of those who took it via the Internet. The question of equivalence arises because there are certain differences between a linear computerized test and a traditional paper-and-pencil test, and also between computerized tests administered via the Internet and those that are not. In the former case, the differences lie in the presentation of the items, the method of answering, how reading comprehension passages and questions with graphic components are presented, and in how time is allotted. Internet-based administration brings other factors into play, for example interruptions to the power supply, non-standard computers in different laboratories, Internet server problems, the impact of heavy traffic on the server, a greater risk of items being compromised, and the challenge of handling problems during the administration itself. The relationship between performance on the experimental test and several background variables (based on a feedback questionnaire) was also examined. The participants were 381 examinees who registered for the October 2008 administration of the Psychometric Test. The paper-and-pencil version was given to 192 of these participants, and 189 were tested via the Internet. Assignment to the two groups was random. Three hundred seventy of the participants in the experiment (185 from each group) took the actual Psychometric Test a month after the experimental administration.

# Conference Abstracts

---

The following conclusions are based on analysis of the results: (1) No significant difference in scores was found between the two groups; (2) No significant differences were found between the scores on the Verbal Reasoning and Quantitative Reasoning sections, but the English scores were significantly higher in the computerized version across all item types; (3) The correlation between the overall experimental scores and scores on the actual test were 0.93 and 0.94 for the computer-based and paper-and-pencil groups respectively; (4) The difference between the two groups in improvement in scores (between the experiment and actual test), both overall and for each section, was not significant; (5) The difference in scores between men and women was the same for both groups; and (6) The correlation between frequency of computer use and performance on the test was similar for both groups. Thus, it was found that the modality of administration, Internet-based or paper-and-pencil, did not affect examinee performance on the Psychometric Test. This holds with respect to item types that we suspected would become more difficult when administered by computer. The results support simultaneous administration in two modalities.

**For further information:** [naomi@nite.org.il](mailto:naomi@nite.org.il)

## Features of a CAT System and Its Application to J-CAT

*Shingo Imai, Y. Akagi, Yamaguchi University, Japan*

*K. Kikuchi, Toho University, S. Ito, TUFS, Japan*

*Y. Nakamura, Tokiwa University, Japan*

*H. Nakasono, Shimane University, Japan*

*A. Honda, APU, and T. Hiramura, TIT, Japan*

A CAT system called J-CAT or Japanese computerized adaptive test, which is operational on the internet or by LAN, has been developed and used as a proficiency test of Japanese at the college level for international students in Japan. We discuss some features of this CAT system, focused on the viewpoint of test administrators. The features discussed in this presentation include registration method, item-pool management, and utilization of test results. We illustrate how this system registers examinees and authenticates them. We also discuss how to manage an item pool; such as uploading items, setting IRT parameters, and setting answering time limits for each item. The system provides useful information for analyzing the results of a test. We highlight some features of a downloadable CSV file of properties of examinees and test results. We show what information is available for an administrator and how an administrator might utilize the information. Examinees also receive feedback on their test results in a report form which is automatically produced at the end of a test. The J-CAT system, which at present contains items for Japanese proficiency, can be also used for tests other than Japanese language if the items are replaced with items of other tests. The system supports Rasch, two-parameter, and three-parameter IRT models.

**For further information:** [imai2002@yamaguchi-u.ac.jp](mailto:imai2002@yamaguchi-u.ac.jp)

## Adaptive Measurement of Cognitive Ability Based on a Person's Zone of Nearest Development

*Marina Chelyshkova and Victor Zvonnikov, State University of Management, Russia*

At the present moment the majority of schools and universities of Russia attach great importance to the cognitive process in education. We think that in modern testing, it is important not only to estimate the degree of knowledge that the person has but also to evaluate cognitive ability, which is more complex than knowledge and skills. The measurement of cognitive ability usually requires special item content, which cognitive learning theories provide. But there are other aspects of such measurement. They are connected with optimization of an item's difficulty and require the application of adaptive testing. We analyzed person characteristic curves

and suggested some methods for adapting the test item's difficulty to the individual. These ideas were combined with the concepts of Russian scientist, L. S. Vigotsky, who suggested the ratio of ability to knowing something (actual zone) to the ability to develop of a person's internal mental forces. His concept allows one to connect the score of actual knowledge with the width of a person's zone of nearest development. We suggested the method for evaluating this connection by using one-parameter and two-parameter models of IRT and expressed it in the form of the system of inequalities, which related the person parameter and item parameters. As applied to measurement of cognitive ability we suggested choosing items that have difficulty appropriate to a person's zone of nearest development instead of traditional scoring approaches in adaptive testing. We developed the connection between the width of the nearest development zone and scores of test items in terms of the difficulty and slope of item characteristic curves. It has allowed us to evaluate a person's cognitive ability and to predict his/her changes of achievement depending on the time factor and the steepness of his/her person characteristic curve. Thus, in such a way we can optimize the difficulty of test items in adaptive testing for measurement of cognitive ability.

**For further information:** [mchelyshkova@mail.ru](mailto:mchelyshkova@mail.ru)

## Implementing Figural Matrix Items in a Computerized Adaptive Testing System: Singapore's Experience

*Poh Hua Tay and Raymond Fong, Ministry of Education, Singapore*

Figural matrix items such as Raven's Standard Progressive Matrices (SPM) are widely used for assessing general intelligence of pupils. Substantial manpower resources are incurred when administering tests on a large-scale basis via paper-and-pencil (P&P). A computer-based test (CBT) would offer the advantages of logistical ease during the data collection stage, and administrative ease during the data entry stage; this is especially so for CAT, as it reduces administration time as well. Unlike P&P and CBT, the most appropriate set of items in a CAT can be adaptively selected for each pupil based on his/her responses to previous items. This permits each pupil to be evaluated on a smaller subset of the total item pool and have a better test experience, as items are chosen based on his/her ability. It also allows the test developer to control the error of measurement to a desired degree of precision.

In this study, an item bank of 195 figural matrix items that are similar to SPM's was created. The psychometric properties of these items were then established after running a trial on a sample of 6,821 Primary 2 pupils (equivalent to Grade 2 pupils who are about eight years in age) of varying academic abilities from 20 coeducational schools in Singapore. Item response theory (IRT) was used to calibrate all the figural matrix items. From this item bank, a P&P prototype, two CAT prototypes (one starts with an easy item, while the other starts with an average item), and a CBT prototype were generated and administered, via the FastTEST Pro v2.3 platform, to four groups of Primary 2 pupils in Singapore. These groups consisted of a total of 948 Primary 2 pupils of varying academic abilities and were selected from 12 coeducational schools. SPM was also administered to all of them via P&P. This project was designed to study the comparability of the abilities of pupils estimated from the different prototypes (P&P, CATs, CBT) and SPM.

**For further information:** [tay\\_poh\\_hua@moe.gov.sg](mailto:tay_poh_hua@moe.gov.sg)

## Constrained Item Selection Using a Stochastically Curtailed SPRT

Jasper T. Wouda and Theo J. H. M. Eggen, CITO, The Netherlands

Computerized classification testing (CCT) can be used to increase efficiency in educational measurement. The truncated sequential probability ratio test (TSPRT) has been widely studied as a decision algorithm in CCT for two or more categories (Spray, 1993; Eggen, 1999). Finkelman (2003) added an algorithm to the TSPRT in the form of stochastic curtailment, to classify an examinee in an even earlier stage of testing. This stochastically curtailed SPRT (SCSPRT) halts testing when a change of classification is possible but unlikely. As can be seen in Finkelman (2003, 2008), the SCSPRT is an extension of the SPRT. It adds stochastic curtailment in the form of two extra stopping rules per level. Stochastic curtailment ceases testing and rejects hypothesis  $H_{0i}$  if given  $k$  observations. The probability that a decision  $D$  will accept  $H_{0i}$ ,  $Pr(D=H_{0i})$ , is not higher than a set value  $1-\gamma$ . It stops testing and accepts  $H_{0i}$  if this probability is at least  $\gamma$ . This method makes use of the suboptimality of the SPRT as used in truncated tests.

In the comparison of performance between the SPRT and SCSPRT (Finkelman, 2003, 2008), results showed a substantial decrease in number of items used per simulee for the SCSPRT, while the percentage of correctly classified simulees remained the same. When using real item parameters and realistic data (Wouda, 2008), this decrease became somewhat smaller, but was still substantial. In order to be applied in real-world tests, however, nonstatistical constraints must also be considered. Different constraints include, for example, content balancing, answer key balancing conflicting items, and item exposure control. In this study, different constraint handling methods will be compared, together with different item selection methods. The applied constraints are content balancing and exposure control. The compared item selection methods will be selection of items at the estimate and selection of items at the cut score. The methods for exposure control that will be compared for the SPRT and SCSPRT are the Sympon-Hetter method, the progressive method, and alpha-stratified testing. The methods for content balancing that will be compared are the Kingsbury and Zara (1989, 1991) approach and the weighted deviation method (WDM) by Stocking and Swanson (1993).

**For further information: [Jasper.Wouda@cito.nl](mailto:Jasper.Wouda@cito.nl)**

## Using Enhanced Effective Response Time to Detect the Extent and Track the Trend of Item Pre-Knowledge on a Large-Scale Computerized Adaptive Assessment

Jie Li and Xiang Bo Wang, ACT, Inc.

In addition to being highly efficient and accurate in terms of scoring, diagnosis, and reporting, CAT is also known for its global ease and reach of test delivery (Wainer et al., 2000; Meijer & Nering, 1999; Parshall, Spray, Kalohn, & Davey, 2002). The latter advantage of CAT, however, also introduces a tenacious problem of potentially exposing items to a high number of examinees due to its high frequency of test administration, which is likely to increase advance or pre-knowledge of items and to jeopardize score validity. Of great concern and interest to the entire educational testing industry is the possibility of validly detecting and tracking the extent that CAT items are exposed. The purpose of this research was (1) to establish population item response times for all items and associated trends for all items with a large-scale international CAT assessment and (2) to investigate the feasibility of applying "effective response time" (ERT; Meijer & Sotaridona, 2006) to detect the extent and track the trend of item pre-knowledge on suspected compromised items on this assessment. The study was based on both operational and simulated data of a large item pool of a large-scale international CAT assessment. This item pool was selected because (1) it had a substantial number of new items that were pretested several years ago when little or no item pre-knowledge could be assumed and (2)

these pretest items had a long history of operational use in subsequent years when item pre-knowledge could have been accumulated. ERT indices for both items and examinee, as described by Meijer & Sotaridona (2006), were computed against a large collection of new items at their pretest time after they passed stringent pretest item quality reviews. The ERT indices from this round were used as null hypothesis benchmarks since no serious item pre-knowledge could be assumed. In addition, simulations were conducted to project the values of these ERT indices, if examinees' response times were reduced by one-half and one-fourth, respectively. Examinees ability estimates on the operational items of this item pool were used for ERT modeling. ERT indices were also computed when all the new items were first used operationally and the results were compared with their pretest counterparts.

**For further information:** [Jie.Li@Act.org](mailto:Jie.Li@Act.org)

## Computerized Adaptive Testing for the Singapore Employability Skills System (ESS)

*Patricia Rickard, CASAS, James B. Olsen, Alpine Testing Solutions, Debalina Ganguli, CASAS, and Richard Ackermann, Team Code, Inc.*

This paper presents and demonstrates innovations in computerized adaptive testing of adult workplace literacy and numeracy skills developed by CASAS and customized for the Singapore Employability Skills System (ESS). The Singapore Workforce Development Agency (WDA) plays a pivotal role in the implementation of the ESS “to enhance the employability and competitiveness of employees and job seekers, thereby building a workforce that meets the changing needs of Singapore’s economy.” CASAS has designed and developed CATs for mathematics, reading, and listening along with computer-delivered tests for writing and speaking that are suitable for adults. The CATs are administered in secure proctored locations using local area networks and an electronic access key (dongle). This paper provides an overview of the project, demonstrations of sample test items from the test battery, a presentation of the test delivery and administration system, a review of test score results and psychometric analyses, and discussion of plans for future enhancements and extensions. The Singapore CATs use the following psychometric procedures: (1) selection of initial item from a random proficiency value near the center of proficiency distribution of the selected item bank, (2) Rasch model calibration and proficiency estimation, and (3) a stopping rule based on a minimum standard error or administration of a specified maximum number of items. Results for the mathematics and reading CATs are presented showing scale score population distributions, stopping rule exit criteria, item exposure distributions, and ability estimate and standard error curves across the item administration sequence. The paper presents summary recommendations for enhancements and extensions with the CAT tests and additional CAT research and validity investigations.

The CAT results are based on examinee samples of approximately 12,000 for the reading tests and 9,000 for the numeracy tests.

**For further information:** [rickard@casas.org](mailto:rickard@casas.org)

## Criterion-Related Validity of an Innovative CAT-Based Personality Measure

*Robert J. Schneider, PDRI; Richard A. McLellan and Tracy M. Kantrowitz, PreVisor, Inc.; Janis S. Houston and Walter C. Borman, PDRI*

This paper blends rigorous and innovative psychometric theory with a practical selection application. We used CAT principles to estimate examinees' personality trait levels through an iterative, IRT-driven, paired-comparison assessment process. The concept has its roots in Thurstone's (1927) Law of Comparative

# Conference Abstracts

---

Judgment. Thurstone conceived of using a paired-comparison procedure to scale stimuli on an interval scale. The idea was that if interval scale personality assessment could be generated with a paired-comparison procedure, then measurement might be made more precise than that yielded by typical Likert-type personality scales, which arguably provide only ordinal level data. Stark and Drasgow (1998) developed an algorithm to implement this process based on Zinnes and Griggs' (1974) probabilistic unfolding model which, in turn, is based on (and extends) the work of Coombs (1950) and Thurstone (1927). Examinees select which of the two statements representing different levels of a personality trait are more descriptive of them, and are then presented with two additional statements, based on their previous selection. Sequences of statement-pairs are selected in a manner that maximizes information in an IRT sense. Statement-pairs are presented for a given personality trait until either (1) a sufficiently low conditional standard error of measurement is reached, or (2) 10 statement-pairs have been presented. This methodology has been used successfully in the Navy (Borman et al., 2001; Houston, Borman, Farmer, & Bearden, 2005). To our knowledge, however, our measure represents the first commercial application of CAT to the personality domain. Our test measures 13 traits selected to represent the broad personality sphere and to be predictive across a wide range of occupations and industries. Our intent was to build in flexibility to create composites of scales relevant to a variety of different work populations to accommodate the differing needs of our clients.

This presentation reports initial validity results. Our CAT personality measure was administered to 1,607 first-line supervisors in eight organizations, each of whom was rated by his/her immediate supervisor. Sample sizes for predictor-criterion pairings ranged from  $n = 745$  to 1,109. To identify a composite of scales relevant to the supervisory position, we conducted a relative weight analysis (Johnson, 2000) to identify the relative importance of each predictor based on its proportionate contribution to  $R^2$ . This procedure controls for multicollinearity among predictors by considering the unique effect of each predictor as well as its effect when combined with the other predictors. Six scales were identified and a weighted sum was computed. The estimated operational validity of the adaptive personality scale composite was .25 against an overall job performance criterion. Graphs showing validity coefficients associated with presentation of different numbers of statement-pairs will also be shown for each scale included in the personality composite, as well as for the composite itself. This information will be very useful in that it will indicate how many statement-pairs must be presented to reach stable (asymptotic) criterion-related validity estimates.

**For further information: [Robert.Schneider@pdri.com](mailto:Robert.Schneider@pdri.com)**

## Computerized Adaptive Testing in Spain: Description, Item Parameter Updating, and Future Trends of eCAT

*Francisco J. Abad and David Aguado, Universidad Autónoma de Madrid*

*Juan Ramón Barrada, Universidad Autónoma de Barcelona*

*Julio Olea, Vicente Ponsoda, and Francisco J. Abad, Universidad Autónoma de Madrid*

eCAT is a computer adaptive testing (CAT) model developed and applied in Spain to assess English proficiency among Spanish speakers. The test was developed by psychometricians from the School of Psychology (Universidad Autónoma de Madrid) and the IIC (Engineering Institute of Knowledge). Psychometricians constructed the item bank and designed the adaptive algorithm. The IIC takes care of the marketing and control of the test delivery via the Internet. At this time, thousands of tests have been administered in the context of the personnel selection processes and for the assessment of undergraduate's language competencies in several Spanish universities. In this presentation we will summarize the work done for the design and updating of the system. We will address four different aspects of eCAT: (1) test construction, including item bank design and calibration, adaptive algorithm, psychometric properties of the scores (reliability and

validity), computerized reports, and software for web-based application; (2) main results of the application (descriptive study of scores, estimation errors, execution time, and exposure rates); (3) analysis of parameter drift and its impact on the scores, assessed by means of a comparison between the estimates of parameters in the initial calibration sample and those obtained under eCAT ordinary operation; and (4) work in progress including item parameter updating, increasing the bank size using online calibration procedures, and calibrating a new bank of items to assess the level of English listening (eCAT-listening).

**For further information:** [fjose.abad@uam.es](mailto:fjose.abad@uam.es)

## Twenty-Two Years of Applying CAT for Admission to Higher Education in Israel

*Naomi Gafni and Yoav Cohen, National Institute for Testing and Evaluation, Israel*

This paper describes the use of CAT in higher education admissions in Israel. This includes: (1) the English-as-a-foreign language (EFL) CAT that has been used by various institutions of higher education for placement purposes for 22 years; and (2) the CAT version of the Psychometric Entrance Test (MIFAM), which has been in use for nine years as a higher education admissions tool for examinees with disabilities. Both applications run in parallel with paper-and-pencil test (PPT) versions. This presentation will focus on the specific procedures used to produce equitable scores across the two media as well as on examining the suitability of the CAT for examinees with disabilities. The paper discusses a host of practical issues that were encountered during conversion of the Psychometric Entrance Test (PET) to a computerized adaptive format. Issues that pertain to the meeting of content specifications, item exposure, item banks, item bank dimensionality, and equating are identified and discussed in the context of evolutionary changes in the MIFAM program.

**For further information:** [naomi@nite.org.il](mailto:naomi@nite.org.il)

## Item Selection and Hypothesis Testing for the Adaptive Measurement of Change

*Matthew Finkelman, Tufts University School of Dental Medicine*

*David J. Weiss, University of Minnesota*

*Gyenam Kim-Kang, Korea Nazarene University*

In a paper presented at the 2007 GMAC CAT Conference, Kim-Kang and Weiss (2007, 2008) described a procedure for the adaptive measurement of change (AMC) for an individual examinee. In this procedure, a CAT is administered at Time 1 to an examinee and the final estimate from that CAT is used to begin a second CAT at Time 2 (a later point in time). The Time 2 CAT continues until the Time 2 95% confidence interval around its estimate does not overlap the Time 1 95% confidence interval. When this occurs “significant change” is said to have occurred for that examinee. Kim-Kang and Weiss compared the performance of the AMC procedure in measuring change with that of change scores from conventional tests based on raw difference scores, residual change scores, and IRT-based difference scores. Their results showed that AMC captured change better than all methods based on conventional tests under a variety of test configurations and levels of true change. They also demonstrated that the AMC procedure was efficient in detecting significant change, requiring an average of 6 to 22 items for different levels of true change.

The present study focused on the detection of change. Two new methods for testing the hypothesis of significant change for a single person were developed and compared to the confidence interval overlap approach. These methods were a likelihood ratio test approach and a Z-test approach. The power and alpha level of these two hypothesis testing methods were evaluated in the context of two CAT item selection methods—Fisher information and a variation of Kullback-Leibler information designed to select items in the



## Item Selection with Biased-Coin Up-and-Down Designs

Yanyan Sheng, Southern Illinois University at Carbondale

A basic ingredient in computerized adaptive testing (CAT) is the item selection procedure that sequentially selects and administers items based on a person's responses to the previously administered items. For decades, maximum information (MI; Lord, 1977; Thissen & Mislevy, 2000) has been widely used as the conventional algorithm for item selection in CAT. This criterion based on Fisher's information, however, only targets the middle difficulty level where a person has about a 0.5 probability of getting the items correct, and hence is not applicable in situations where a different percentile is desired. In addition, MI heavily relies on an accurate estimation procedure that works well in all testing situations. Nonetheless, studies have shown that such a procedure is not readily available.

The biased-coin up-and-down design (BCD; Durham & Flournoy, 1994) has been widely used in bioassay for sequential dosage level selection because it can target any arbitrary percentile in addition to being efficient (Bortet & Giovagnoli, 2005). Since the problem in bioassay shares many similarities with CAT, it is reasonable to believe that the item selection algorithm based on the BCD, which does not rely on an accurate trait estimate in every step of CAT administrations, provides an efficient alternative to (while being more flexible than) the conventional method. The development of this selection algorithm is essential as schools, professional organizations, and private companies seek to make CAT flexible enough to be implemented in wider testing applications.

The purpose of this study was to illustrate the use of the BCD in CAT and further evaluate its utility by comparing it with the conventional MI algorithm. For ease of comparisons, this study focused on the one-parameter item response function. To investigate the utility of the BCD in CAT, two *Monte Carlo* simulation studies were conducted, using either a fixed- or a random- stopping rule. With fixed-stopping rule, the number of items administered was manipulated ( $k = 5, 10, 30, 100$ ) and the item pool was fixed to have 100 different difficulty levels, whereas with random-stopping rule, the number of different difficulty levels in the item pool was manipulated ( $n = 10, 30, 50, 100$ ). In either case, CAT responses were simulated for persons whose actual trait levels were 0 (average), -1 (1 standard deviation below the average), and -2 (2 standard deviations below the average), and the target difficulty level was at the 20th, 50th or 80th percentile. Each adaptive testing simulation began the trait estimation with an initial value of 0 and proceeded with the maximum likelihood method. The results suggested that item selection with the BCD is more flexible in targeting any arbitrary percentile of the difficulty levels. With respect to the accuracy of the trait estimation, MI performs slightly better with fixed-stopping rule, whereas the BCD is considerably better for tests with a small number of difficulty levels or persons whose trait levels are not at the extremes with random-stopping rule.

**For further information:** [ysheng@siu.edu](mailto:ysheng@siu.edu)

## A Burdened CAT: Incorporating Response Burden with Maximum Fisher Information for Item Selection

Richard J. Swartz, The University of Texas M. D. Anderson Cancer Center  
Seung W. Choi, Northwestern University Feinberg School of Medicine

Widely used in various educational and vocational assessment applications, CAT has recently begun to infiltrate the patient-reported outcomes (PRO) arena. Several differences exist between PRO-CAT and "achievement CAT." Polytomous, rather than binary, items are more appropriate for PROs; constructs are often quasi-traits with skewed distributions; informative items cannot always be generated along the important range of the

trait; and in many patient populations conditions exist so that patients cannot tolerate longer tests. Reducing this response burden has been one of the main reasons for consideration of CAT in the PRO arena. Although successful in reducing burden, many of the current CAT algorithms do not formally consider patient or examinee burden as part of the item selection process. In the PRO setting, many CAT applications simply limit the maximum number of items to be administered. This study uses a loss function approach motivated by decision theory to develop an item selection method that incorporates burden into the Maximum Fisher's Information (MFI) item selection method.

We compared several different loss functions representing varying degrees of burden, including a no-burden condition as a baseline. An item bank of 62 polytomous items measuring depressive symptoms was used to compare the different methods. The items were calibrated with the graded response model using 730 patients and caregivers from the M. D. Anderson Cancer Center. For each condition, we used two different response datasets to simulate CAT instruments. One dataset consisted of the real responses from the 730 patients and caregivers who answered all the items. The second dataset consisted of simulated responses to all the items based on a grid of values with replicates at each grid point. The MFI-burden algorithm for item selection results in tests that are on average shorter (depending on the degree of burden) than those obtained using MFI alone, but without severely affecting the standard error of measurement. In particular the loss function incorporating burden protects respondents from receiving longer tests when their estimated trait score falls in a location where there are few informative items. This is very useful in PRO assessment where burden to the patient is a concern.

**For further information: [rswartz@mdanderson.org](mailto:rswartz@mdanderson.org)**

## Adaptive Item Calibration: A Simple Process for Estimating Item Parameters Within a Computerized Adaptive Test

*G. Gage Kingsbury, Northwest Evaluation Association*

The characteristics of computer-adaptive testing (CAT) change the characteristics of the field testing that is necessary to add items to an existing measurement scale. The process used to add field test items to a CAT might lead to scale drift (van der Linden & Glass, 2000; Ban et al, 2001). In addition to this measurement concern, adding randomly chosen field test items to a test might disrupt the performance of an examinee by administering items of inappropriate difficulty. The current study makes use of the transitivity of examinee and item in IRT to describe a process for adaptive item calibration. In this process an item is successively administered to examinees whose ability levels match the performance of a given field test item. By treating the item as if it were taken in an adaptive test, examinees can be selected who provide the most information about the item at its momentary difficulty level. Throughout the calibration process, the momentary difficulty estimate is updated and used in the process of item selection for all examinees. The item calibration can be completed when a fixed number of examinees have seen the item of interest, or when the momentary difficulty level for the item stabilizes to a predetermined variability. This approach should provide a more efficient procedure for estimating item parameters. While the procedure is not specifically designed to create an optimal calibration sample in the manner described by Holman and Berger (2001), it should result in the item being administered to a set of individuals that more closely approximates optimality.

The process is described in detail within the context of the one-parameter logistic IRT model. The process is then simulated using 10 replications of the calibration of 100 items to identify whether it produces more accurate and efficient item parameter estimates than random presentation of field test items to examinees.

Results indicate that adaptive item calibration is more accurate for small sample sizes. With additional research, adaptive item calibration might provide a viable approach to expanding item pools in settings with small sample sizes or settings with a need for large numbers of items.

**For further information:** [gage.kingsbury@nwea.org](mailto:gage.kingsbury@nwea.org)

## On-the-Fly Item Calibration in Low-Stake CAT Procedures

*Sharon Klinkenberg, Marthe Straatemeier, and Han van der Maas, University of Amsterdam*

We present a new model for computerized adaptive progress-monitoring. This model is used in the Math Garden, a web-based monitoring system, which includes a challenging web environment for children to practice arithmetic skills. The Math Garden is a CAT web application, which tracks both accuracy and response time. Using a new model (Maris, in preparation) based on the Elo (1978) rating system and an explicit scoring rule, estimates of ability level and item difficulty are updated every trial. Items are sampled with a mean success probability of .75, making the tasks challenging yet not too difficult. By integrating the response time in the scoring rule, we try to compensate for the loss of information associated with the high success rates (van der Maas & Wagenmakers, 2005). In a period of eight months, our sample of 1,053 children completed more than 850,000 arithmetic problems. The children completed about 25% of these problems outside their school hours. Results show good validity and reliability, high pupil satisfaction measured in playing frequency, and good diagnostic properties. The ability scores correlated highly with the Dutch norm-referenced general math ability scale of the pupil monitoring systems of CITO. Also, test-retest reliability analysis showed high correlations. In view of the satisfactory validity and reliability of the person ability estimators, our method opens the door to on-the-fly item calibration in low-stakes testing.

**For further information:** [S.Klinkenberg@uva.nl](mailto:S.Klinkenberg@uva.nl)

## An Automatic Online Calibration Design in Adaptive Testing

*Guido Makransky and Cees. A. W. Glas, University of Twente, The Netherlands*

An accurately calibrated item bank is essential for a valid CAT. In some settings, however, such as occupational testing, there is limited access to examinees for such calibration. As a result of the limited access to possible examinees, collecting data to accurately calibrate an item bank in an occupational setting is usually difficult. In such a situation, the item bank can be calibrated online in an operational setting. This study explores three possible automatic online calibration strategies with the intent of calibrating items accurately while estimating ability precisely and fairly. That is, the item bank is calibrated in a situation where examinees' ability is assessed throughout the calibration design. The three calibration strategies represent a sample of possible designs on a continuum ranging from one extreme, where items are calibrated at a single point in time, to the other extreme where items are calibrated constantly after each exposure. A simulation study was used to identify the optimal calibration strategy. The outcome measure was the mean absolute error of the ability estimates of the examinees participating in the calibration phase. Manipulated variables included the calibration strategy, the size of the calibration sample, the item response mode, and the size of the item bank. The results of the study give an overview of the benefits of each strategy for different applied conditions, and provide viable calibration design options for test development companies that find it difficult to get examinees in the development phases of a test.

**For further information:** [guidomakransky@gmail.com](mailto:guidomakransky@gmail.com)

## Investigating Cheating Effects on the Conditional Sympon and Hetter Online Procedure with Freeze Control for Testlet-Based Items

Ya-Hui Su, University of California, Berkeley

In CAT, if a group of examinees purposefully memorize items and distribute them to other prospective examinees, it certainly ruins the equality and accuracy of CAT. Steffen and Mills (1999) investigated this effect and found that the more compromised the items and the more effective the cheating, the more severe the overestimation for the recipients, especially for those with low ability levels. Su, Chen, and Wang (2004), pointed out that the overestimation for the recipients was more severe when the sources had diverse ability levels, because more items were compromised. Su and Wang (2007) proposed an item exposure control procedure, called the conditional Sympon and Hetter (Sympon & Hetter, 1985) online procedure with freeze control (denoted as SHCOF) procedure. Results showed it superior to many other conventional procedures in terms of measurement and operational efficiency. To assess the cheating effect, Su and Wang (2008) used the SHCOF procedure in a CAT, and found it could obtain precise estimation for persons in real time without requiring simulations to generate item exposure under a unidimensional context. In the past, little research has been done to investigate cheating effects within a testlet context. Hence, it is of great value to ascertain whether the SHCOF is also less affected by the cheating between examinees under a testlet context, when compared to a popular procedure such as the conditional multinomial method (SLC; Stocking & Lewis, 1998). The goal of this study was to use simulations to investigate how these two item exposure control procedures would perform under various cheating conditions. It was hypothesized that SHCOF would be less affected by cheating than SLC.

Four independent variables were manipulated: (1) ability level of sources, (2) ability distribution of recipients, (3) cheating conditions (no cheating, inefficient cheating, efficient cheating, and perfect cheating), and (4) item exposure control procedure (SHCOF and SLC). The root mean squared error (RMSE) was computed to describe the cheating effects; the more serious the cheating effect, the larger the RMSE. Under the no-cheating condition, there is no significant difference in RMSE between SHCOF and SLC. It was also found that SLC had more serious inflation on RMSE than SHCOF under the perfect cheating condition. As the cheating condition got more severe, the overestimation for the recipients got more severe when the SLC was used. In addition, the more diverse the ability of the sources, the larger the RMSE and the mean positive bias would be. More importantly, SHCOF had smaller RMSE than SLC. This was because only SHCOF could simultaneously monitor item exposure and test overlap rates online. SHCOF could obtain precise estimation for persons without requiring simulations to generate item exposure before using in an operational CAT. If test items are memorized by sources and shared with recipients, CAT becomes unfair because the ability levels of the recipients will be overestimated. In this study, it was found that SHCOF was less affected by cheating than SLC. Hence, the SHCOF procedure can be safely implemented in operational CAT.

**For further information: [yahuisu@berkeley.edu](mailto:yahuisu@berkeley.edu)**

## The Nine Lives of CAT-ASVAB: Innovations and Revelations

Mary Pommerich, Daniel O. Segall, and Kathleen E. Moreno, Defense Manpower Data Center

The Armed Services Vocational Aptitude Battery (ASVAB) is administered annually to more than one million military applicants and high school students. ASVAB scores are used to determine enlistment eligibility, assign applicants to military occupational specialties, and aid students in career exploration. The ASVAB is administered as both a paper-and-pencil (P&P) test and a CAT. CAT-ASVAB holds the distinction of being the first large-scale computer adaptive test battery to be administered in a high-stakes setting. Approximately

two-thirds of military applicants currently take CAT-ASVAB; long-term plans are to replace P&P-ASVAB with CAT-ASVAB at all test sites. Given CAT-ASVAB's pedigree—approximately 20 years in development and 20 years in operational administration—much can be learned from revisiting some of the major highlights of CAT-ASVAB history. This paper traces the progression of CAT-ASVAB through nine major phases of development including research and development of the CAT-ASVAB prototype, the initial development of psychometric procedures and item pools, initial and full-scale operational implementation, the introduction of new item pools, the introduction of Windows administration, the introduction of Internet administration, and research and development of the next generation CAT-ASVAB. A background and history is provided for each phase, including discussions of major research and operational issues, innovative approaches and practices, and lessons learned.

## The CAT-DI Project: Development of a Comprehensive CAT-Based Instrument for Measuring Depression

*Robert D. Gibbons, University of Illinois at Chicago*

The combination of IRT and CAT has proven invaluable in educational measurement. More recently, enormous reduction in patient and physician burden have been demonstrated using IRT based CAT in the area of mental health measurement problems (Gibbons et al., 2008). CAT administration of a 626-item mood and anxiety spectrum disorder inventory revealed that an average of 24 items per examinee were required to provide impairment estimates with a correlation of 0.93 with the original complete scale. Furthermore, the CAT-based scores revealed twice the effect size than the total scale score in terms of differentiating patients with bipolar disorder based on the mood disorder subscale, despite an 83% reduction in the average number of items administered. These preliminary findings led to further interest and funding by the National Institute of Mental Health to develop a CAT-based instrument for the screening of major depressive disorder (CAT Depression Inventory—CAT-DI) that can be used for routine screening of depression in general medical practice settings as well as specialty mental health clinics. A recent supplement to the parent CAT-DI grant extends our work on CAT for mental health measurement to CAT for diagnostic assessment of depression and other psychiatric disorders. The CAT Major Depressive Disorder (CAT-MDD) project will explore four different statistical/psychometric models for estimating the probability of an underlying discrete major depressive disorder based on self-administered symptom ratings that are adaptively administered. The ultimate objective of this program of research is to reduce patient and physician burden in terms of screening and diagnosing depression in general practice settings. Potential benefits include reduction in health care costs produced by high rates of service utilization among patients with an undiagnosed depressive illness, increased detection of depressive disorders, and increased access to quality mental health care for patients in need of such services.

**For further information: [rdgib@uic.edu](mailto:rdgib@uic.edu)**

## Development of a CAT to Measure Dimensions of Personality Disorder: The CAT-PD Project

*Leonard J. Simms, University of Buffalo*

This presentation describes the CAT-PD project, a funded, multi-year study designed to develop an integrative and comprehensive model and measure of personality disorder trait dimensions. Our general study aims are to (1) identify a comprehensive and integrative set of dimensions relevant to personality pathology, and (2) develop an efficient CAT method—the CAT-PD—to measure these dimensions. To accomplish our general

# Conference Abstracts

---

goals, we plan a five-phase project to develop and validate the model and measure. The presentation describes the project generally, the results of Phase I (which is focused on content domains and initial item bank development), and our plans for IRT/CAT with these item banks. In particular, I will focus on how the item banks will be used, the possible IRT models we are considering for item bank calibration, the CAT algorithms we are planning to test, and our methods for deciding on a final set of procedures for the completed CAT-PD measure. Finally, I will discuss the CAT and IRT challenges that we anticipate facing in the future.

**For further information:** [lsimms@buffalo.edu](mailto:lsimms@buffalo.edu)

## MEDPRO Project: An SBIR Project for a Comprehensive IRT and CAT Software System

### Part I: The IRT Software

*David Thissen, The University of North Carolina at Chapel Hill and Scientific Software International*

The IRTPRO (Item Response Theory for Patient-Reported Outcomes) component of the MEDPRO Project is an entirely new application for item calibration and test scoring using IRT. A Fall 2009 release of this software is anticipated. This presentation briefly describes its features, user interface, and output. IRTPRO provides maximum likelihood calibration of items fitted with the 1PL, 2PL, 3PL, Graded, Generalized Partial Credit, and Nominal IRT models in any combination, using one of three estimation algorithms: (1) Bock-Aitkin EM, (2) adaptive quadrature, or (3) Metropolis-Hastings Robbins-Monro (MHRM). Unidimensional or multidimensional IRT models might be used; among multidimensional models, the implementation performs full-information estimation for exploratory and confirmatory models, including the special-case treatment appropriate for bifactor models. Analysis of differential item functioning (DIF) is also provided, using the Wald test, with accurate item parameter error variance-covariance matrices computed using the Supplemented EM (SEM) algorithm. Several goodness-of-fit and diagnostic statistics are reported. Standard *maximum a posteriori* (MAP) and *expected a posteriori* (EAP) estimates of the latent variable(s) for item response patterns might be computed, as well as (weighted) summed-score to scale score translation tables.

**For further information:** [dthissen@email.unc.edu](mailto:dthissen@email.unc.edu)

### Part II: The CAT Software

*Nathan A. Thompson, Assessment Systems Corporation*

The CAT software for MEDPRO is designed to provide a comprehensive environment for the design and delivery of CATs. It consists of two main components: CATSIM and FASTCAT, in a package called CATPRO (Computerized Adaptive Testing for Patient-Reported Outcomes), which will be designed to interface with IRTPRO. CATSIM will be a major expansion of Assessment Systems' (ASC) POSTSIM software. CATSIM will implement post-hoc simulations, *Monte Carlo* simulations, and hybrid simulations of CATs. New features in CATSIM will include the addition of CAT for polytomous IRT models, item selection constraints (content balancing item exposure controls and "enemy" items), and an expanded set of termination options. FASTCAT will be an expansion of ASC's FastTEST Professional Testing System that includes all the options in CATSIM applied to the delivery of live CATs in a Windows environment. Output from both CATSIM and FASTCAT will optionally be available in formats directly importable into IRTPRO for analysis and the parameter output from IRTPRO will be directly importable into both CATSIM and FASTCAT.

**For further information:** [nthompson@assess.com](mailto:nthompson@assess.com)

## Reviewing Test Overlap Rate and Item Exposure Rate as Indicators of Test Security in CATs

*Juan Ramón Barrada and Julio Olea, Universidad Autónoma de Barcelona, Spain*

*Vicente Ponsoda, Universidad Autónoma de Madrid, Spain*

*Francisco J. Abad, Universidad Autónoma de Madrid, Spain*

Test security is a major concern in CAT because of the possibility of item sharing among examinees. A CAT will be considered more secure the lower the overestimation of the examinee's trait level is due to item pre-knowledge. The common measures of test security have been the overlap rate between examinees and the distribution of item exposure rates. Usually, these indicators of test security have been evaluated when no item disclosure is present. We justify that lower overlap rates or less skewed distributions of usage of the items might not lead to safer CATs. The main ways of increasing security are to reduce: (1) the probability of item pre-knowledge of the first items administered, and (2) the overlap rate for high trait levels. In these conditions, there would be many different routes to obtain a high trait level estimation and it would be difficult for an examinee with item pre-knowledge to incorporate one of these routes. Progressive and proportional methods offer these characteristics. We show that these two methods are safer than the alpha-stratified method, a method with a much lower overlap rate. In fact, when the alpha-stratified method is applied, there is a "golden source of information." An examinee with high trait level sharing item content is the best option for increasing trait estimation. When the progressive or proportional methods are applied, there is no source of information that fits to all the possible recipients. With these two methods, recipients and sources should have a similar trait level to lead to an important increment of trait estimation.

**For further information:** [juanramon.barrada@uab.es](mailto:juanramon.barrada@uab.es)

## Optimizing Item Exposure Control and Test Termination Algorithm Pairings for Polytomous Computerized Adaptive Tests with Restricted Item Banks

*Michael Chajewski and Charles Lewis, Fordham University*

Much of the item response theory (IRT) and item exposure control literature regarding CAT has focused on the assessment of the impact of exposure control algorithms on frequency of item use, estimation precision, test bias, and overlap as well as item pool utilization and observed root mean square error rates. Most inquiries into these pertinent issues, however, have limited their inquiries to fairly large educational assessment-based item bank situations, which are less common in other areas into which CAT has been expanding. This paper discusses the results of a simulation study that focused on the pairing of item exposure control algorithms and test termination criteria within the specific framework of polytomous CATs using restricted item banks. Based on prior comparative and exploratory research by Chang and Twu (1998), Revuelta and Ponsoda (1998), Pastor, Dodd and Chang (2002), French and Thompson (2003), Davis (2002, 2004), Davis and Dodd (2005), Barada, Mazuelo and Olea (2006), Georgiadou, Triantafyllou, and Economides (2007), and Barada, Olea and Abad (2008), six item exposure control algorithms and four test termination criteria were selected. Item exposure controls included the progressive-restricted maximum information method, Stocking and Lewis conditioning on estimated ability, target exposure control (TEC), Sympon-Hetter conditional strategy (SHC), 0-1 -stratified strategy (0-1STR), and the combined -stratified Sympon-Hetter method (STR-SH). The impact of these six algorithms was evaluated in their optimization of small item bank adaptive instruments using fixed length or fixed standard error (or Fisher target information) test termination criteria. Just like educational large test item



## Comparison of Adaptive Bayesian Estimation and Weighted Bayesian Estimation in Multidimensional Computerized Adaptive Testing

Po-Hsi Chen, Taiwan Normal University

The goal of the research was to compare two new Bayesian estimation methods—the adaptive Bayesian estimation and weighted Bayesian estimation—in multidimensional computerized adaptive testing (MCAT). *Monte Carlo* simulation and a multidimensional item response model, the multidimensional random coefficients multi-nominal logit model (Wang, Wilson, & Adams, 1997), were used in this research. Ten to sixty items of two-dimensional CAT were used with adaptive Bayesian, weighted Bayesian, and traditional Bayesian estimation. The dependent variables were conditional bias and the root mean square error (RMSE). Results indicated that these two new Bayesian approaches resulted in less regression bias than the traditional Bayesian estimation; however, weighted Bayesian estimation was more stable than the adaptive Bayesian estimation. The applications and suggestions for use of weighted Bayesian estimation are addressed.

**For further information:** [chenph@ntnu.edu.tw](mailto:chenph@ntnu.edu.tw)

## Comparison of Ability Estimation and Item Selection Methods in Multidimensional Computerized Adaptive Testing

Qi Diao and Mark Reckase, Michigan State University

The impetus for this research is the lack of guidelines for designing multidimensional computerized adaptive tests (MCATs). There has been some research on unidimensional CAT on the properties of ability estimation and item selection methods (e.g. Weiss & McBride, 1984; van der Linden & Pashley, 2000). In the literature on MCAT, however, most studies use a single ability estimation and item selection method because they focus on other aspects of adaptive testing (e.g. Li Ip & Fuh, 2008). The only study on a comparison of different ability estimation and item selection methods for MCAT is Tam (1992). But that was before most currently used methods (e.g. Segall, 1996; Veldkamp & van der Linden, 2002) were developed. Also, most of the research has used two-dimensional cases, but we believe at least three dimensions are needed. In the proposed study, three ability estimation methods were compared. The first is the general maximum likelihood method (Segall, 1996). A problem when maximum likelihood is used is that estimates of location are not finite when the number of test items is small. One solution offered in Reckase (2009) is fixed-step-size maximum likelihood. This method updates the estimates of ability location with a fixed increment when infinite estimates are encountered. The third method is Bayesian estimation (Segall, 1996).

In the proposed study, four item selection methods were compared. The first is maximizing the determinant of the Fisher information matrix (Segall, 1996). The second is minimizing the trace of the inverse of Fisher information matrix (Mulder & van der Linden, 2008). The third is maximizing the decrement in the volume of the Bayesian credibility ellipsoid (Segall, 1996). The last is maximizing the Kullback-Leibler information (Veldkamp & van der Linden, 2002). The ability estimation and item selection methods conditioning were compared using different priors and test length. The item pool was simulated based on data from the Michigan Educational Assessment Program mathematics test for seventh graders. Mean bias and mean squared error (MSE) were used as a measure of estimation precision. Test lengths of 20 and 50 were generated and results were compared. For testing the impact of priors on the Bayesian method, a multivariate normal distribution with mean 0 and an identity variance-covariance matrix as in the real MEAP 2005 data were used and final ability estimates were compared. The maximum likelihood estimation method did not perform well for the test length of 20. When test length was 50, the estimates were much better. The fixed-step-size maximum likelihood method fixed the problem of estimates not converging and the results were comparable to the Bayesian

method. Bayesian estimates were regressed toward 0 because Bayesian estimates tend to be statistically biased toward the mean of the prior. The standard errors of the estimation were smaller than the maximum likelihood method. Maximizing the determinant of the Fisher information matrix and minimizing the trace of the inverse of Fisher information matrix were comparable. When Bayesian ability estimation was used, the performance of Kullback-Leibler information was slightly better than the Bayesian item selection method with the test length 20. Those two methods were comparable with a test length of 50.

**For further information:** [diaoqi@msu.edu](mailto:diaoqi@msu.edu)

## Multidimensional Adaptive Test: The Application of Kullback-Leibler Information

*Chun Wang and Hua-Hua Chang, University of Illinois at Urbana-Champaign*

In adaptive testing, items are selected sequentially to match the updated ability of the examinee. Numerous item selection algorithms for item pools calibrated under unidimensional IRT models have been well developed. The assumption of unidimensionality can be easily violated, however, especially when the test covers broad content areas. In the presence of multidimensionality, instead of obtaining  $m$  separate unidimensional ability estimates, multidimensional IRT (MIRT) that provides an  $m$ -dimensional vector estimate might be a better choice. Previous researchers have shown that this kind of simultaneous estimation of abilities from different dimensions yields more accurate estimates, since it takes into account the correlational structure of those abilities. Built on MIRT, multidimensional adaptive testing (MAT) can, in principle, provide a promising choice in ensuring efficient estimation of each ability dimension. Currently, two item selection procedures have been developed for MAT, one based on Fisher Information embedded within a Bayesian framework, and the other using Kullback-Leibler (KL) information. Since Fisher information extends to a matrix, instead of a single value in multidimensional ability space, item and test information are no longer independent of each other. Therefore, the nice additive property of Fisher information does not apply to MAT. Alternatively, Kullback-Leibler information remains a single value and thus keeps its additive property.

It is well-known that in unidimensional IRT, the second derivative of KL information (also termed “global information”) is Fisher information evaluated at  $\theta_0$ . This paper first generalizes the relationship between these two types of information in two ways—the analytical result is given as well as the graphical representation to enhance interpretation and understanding. It is shown that the complete Fisher information matrix can be easily recovered from KL information, and the diagonals of the matrix equate to the curvature of the KL information curve, evaluated with respect to each dimension separately. Second, a KL information index is constructed in MAT, which represents the integration of KL information over all of the ability dimensions. In geometric interpretation, this index is analogous to the volume under the information surface when only two dimensions are considered. This paper further discusses how this index correlates with the item discrimination parameters. In the two-dimensional case, an analytical derivation shows that the size of the KL information index depends largely upon the sum of the squared item discrimination parameters, which is also termed “multidimensional discrimination.” The results would lay a foundation for future development of item selection methods in MAT which can help equalize the item exposure rate. Finally, a simulation study will be conducted to verify the above results. The connection between the item parameters, item KL information, and item exposure rate is demonstrated for an empirical MAT delivered by an item pool calibrated under two-dimensional IRT.

**For further information:** [cwang49@illinois.edu](mailto:cwang49@illinois.edu)

## Multidimensional Adaptive Personality Assessment: A Real-Data Confirmation

Alan D. Mead, Avi Fleischer, and Jessica D. Sargent, Illinois Institute of Technology

Although CAT was developed in the context of ability tests (Weiss, 1982), studies have since demonstrated the effectiveness of CAT for measuring attitudes and personality. For example, Koch, Dodd, and Fitzpatrick (1990) applied the rating scale model to a Likert-scale attitudinal questionnaire. The rating scale model (an extension of the one-parameter logistic model for polytomous data) was found to fit the data very well and, although they noted item pool issues, succeeded in measuring effectively. Other studies have found similar results for personality assessments, suggesting that perhaps half the items of an assessment are needed to achieve comparable reliabilities (Waller & Reise, 1989; Reise & Henson, 2000). One issue that has not been extensively treated in prior literature, however, is the multidimensional nature of most personality assessments. Prior research has generally applied unidimensional CAT to individual scales. Segall (1996) presented a multidimensional CAT (MCAT) methodology where correlations between the factors could be leveraged to administer and score items even more efficiently. Mead, Segall, Williams, and Levine (1997) described a *Monte Carlo* simulation of the adaptive administration of the 16PF Questionnaire (Cattell, Cattell, & Cattell, 1993; Conn & Rieke, 1994) using Segall's MCAT method. As in Segall's simulation, the MCAT method was effective in allowing additional reductions in assessment length, beyond those typically encountered with unidimensional CAT. For example, overall assessment length could easily be cut in half with small decrements in scale reliabilities.

The purpose of the current study was to extend the results of the *Monte Carlo* simulation (Mead, et al, 1997) to real data. This study is important for two reasons. First, it is always important to show that simulated results generalize to actual use. More importantly, recent research on personality (research that specifically included the 16PF; Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001) has suggested that traditional IRT models do not fit personality data well and might not be the most appropriate models (Stark, Chernyshenko, Drasgow, & Williams, 2006). If the IRT model is a poor fit to 16PF data, the *Monte Carlo* results will not hold for real data. On the other hand, if the real-data results replicate the simulation results, then we might assume that traditional IRT models fit 16PF data sufficiently well. We obtained archival data from the administration of the 16PF Questionnaire to approximately 5,000 individuals and the two-parameter logistic model was fit to the items using BILOG-MG 3.0. Segall's (1996) software was adapted to read the actual responses of the individuals for a real-data simulation. Results generally supported the use of MCAT with 16PF items. Correlations between actual 16PF scores and MCAT trait estimates were high (averaging .91 to .82) for MCAT tests shortened by up to 40% to 50% while shorter MCAT tests had moderate correlations (averaging .72 to .58). The presentation will also discuss results for the pool usage (about a third of the pool had exposures greater than 90%), efficiency for individuals with extreme scores, and practical considerations for adaptive personality assessment.

**For further information:** [jsergent@iit.edu](mailto:jsergent@iit.edu)



## Adaptive Computer-Based Tasks Under an Assessment Engineering Paradigm

*Richard M. Luecht, The University of North Carolina at Greensboro*

Assessment engineering (AE; Luecht, 2007, 2008a, 2008b; Luecht, Gierl, Tan, and Huff, 2006) is a highly structured way of designing constructs and building instruments and associated scales that measure those constructs. By using construct maps, evidence models, task models, and templates, AE makes it possible to generate extremely large numbers of test forms with prescribed psychometric characteristics (e.g., targeted measurement precision). This paper presents an extension of AE to include computerized-adaptive performance tasks (CAPTs). In a traditional CAT, each item is selected to maximize the measurement precision relative to a provisional estimate of some latent trait. CAT requires every item to be calibrated using an appropriate IRT model so that estimates of item difficulty (location) and other characteristics can be used in the item selection process. Under AE, task models and templates can generate large classes of items. In turn, individual items inherit the estimated psychometric characteristics of the task models and/or templates. A hierarchical Bayesian framework is used for calibration and to quantify uncertainty associated with the class of items sharing estimated item parameters (cf. Glas and van der Linden, 2003). With CAPTs, features or components of the task models and/or templates are altered in real-time to actually vary the task difficulty in a systematic way. By applying a maximum information criteria to an item generation algorithm scripted as part of an AE template, the task features can be selected to create highly variable computer-based performance tasks (i.e., items) that effectively adapt themselves to the proficiency of the examinee. In this sense, the ensuing performance task or items become semi-intelligent measurement agents. The theoretical foundations for CAPTs will be presented in the context of several measurement scenarios. This paper will also present the hierarchical Bayes calibration framework and algorithms for item generation.

**For further information: Email: [rmluecht@uncg.edu](mailto:rmluecht@uncg.edu)**

## Security and Memorability of Innovative Items

*Anne Wendt, National Council of State Boards of Nursing*

*J. Christine Harmes, James Madison University*

Innovative item types have been used in CAT for several years. It has been suggested that innovative items might be more memorable than traditional text-based items; thus potentially posing security concerns. Given that the production of innovative items tends to be expensive and time-consuming, it is especially important that examination programs carefully consider security issues when introducing innovative item formats. This research investigated the degree to which examinees remember various types of items by developing two 70-item tests comprised of paired items (e.g., an innovative item on Form A paired with a text-based conversion of the innovative item on Form B). A five-point (0-4) rating scale was developed to categorize the amount of information remembered by examinees. The average ratings of items were 0 to 1.49 with the average rating of examinee remembrances of items being 1 (which corresponded to having a general impression of the item). Additionally, a qualitative framework for the type of information remembered was developed. In general, examinees primarily remembered general information about the questions, item formats, scenarios, and stimulus components and topics. In very few cases was enough detail provided about the distractors, specific clients or laboratory values to compromise an item.

**For further information: [awendt@ncsbn.org](mailto:awendt@ncsbn.org)**

## Evaluation of a Hybrid Simulation Procedure for the Development of Computerized Adaptive Tests

Steven W. Nydick and David J. Weiss, University of Minnesota

The ideal CAT has a large item bank with a wide range of item difficulties; furthermore, in order for the test to provide equiprecise measurements, there must be items that provide sufficient information across the full range of (Weiss, 1982). Post-hoc simulations have been proposed as a means of fine-tuning a CAT for live administration; indeed, Gibbons, Weiss et al. (2008) demonstrated that the results of post-hoc simulations well predict the outcomes of a live CAT. Before examining CAT test characteristics (e.g., SEM) with a post-hoc CAT simulation, however, each examinee must have provided a response to each item in a bank. But if the item bank is very large (e.g., 1,000), it might not be reasonable to expect any examinee to respond to all the items without factors external to the trait (e.g., fatigue) affecting his/her score. Frequently, because they tend to be large, CAT item banks are calibrated using concurrent calibration methods, which estimate IRT parameters from an incomplete data matrix including a set of linking items (e.g., Kim & Cohen, 1998). This paper proposes and evaluates the performance of a hybrid simulation procedure for use in developing CATs that employs these sparse, concurrent-linking matrices. The hybrid procedure estimates  $\theta$  for each examinee with the item parameters estimated from the sparse linking matrix in conjunction with the set of item responses for each examinee. Then, the  $\theta$  estimate for each examinee is used with *Monte Carlo* simulation methods to impute the examinee's missing data, resulting in a complete response vector for each examinee—part real-item responses and part imputed simulated data. A post-hoc simulation is then implemented with the hybrid response matrix.

Two IRT models were used—two- and three-parameter logistic. From a simulated data matrix of 620 items and 1,000 examinees, either two, four, five, or ten item/examinee blocks were selected, with 20 anchor items and the remainder of the items and simulees divided randomly into groups. Then, responses to items not belonging to a simulee's group were deleted, resulting in data matrices with anywhere from 49% to 87% missing data. Parameters were estimated for both the matrix of full responses and the matrix of partial responses and  $\theta$  was estimated for each simulee. The new estimates of  $\theta$  and the estimated IRT parameters were then used to simulate new responses. POSTSIM (Assessment Systems Corporation, 2007) performed a fixed termination (40 items) and a variable termination (SEM  $\leq .20$ ) post-hoc CAT on each matrix. For both the fixed and variable termination criteria, the hybrid CAT with parameters estimated from the full matrix of responses (HFP) had accuracy close to that of the hybrid CAT with parameters estimated from the partial matrix of responses (HPP), yet it also had efficiency close to that of a CAT performed on the full matrix of responses (FFP). The HPP had correlations with the FPP full-test well into the .90s; HPP and FPP performed poorly only near the limits of estimating the 3PL (80 items per group). These results suggest that meaningful hybrid simulations can be performed with sparse data matrices involving up to almost 80% missing/imputed data. The simulation results were replicated with a real data set.

**For further information: [nydic001@umn.edu](mailto:nydic001@umn.edu)**

## Computerized Adaptive Testing for Cognitive Diagnosis

*Ying Cheng, University of Notre Dame*

Computerized adaptive testing (CAT) is a new mode of testing that enables more efficient and accurate recovery of latent traits. Traditionally, CAT is built upon IRT models that assume unidimensionality. With the advances of latent class models (LCM) and an increasing number of applications of them in testing and measurement, an interesting question that arises is how to build a CAT based on a LCM. Tatsuoka (2002) and Tatsuoka and Ferguson (2003) established a general theorem on the asymptotically optimal sequential selection of experiments to classify finite, partially ordered sets. Xu, Chang and Douglas (2003) proposed two heuristics on the basis of Tatsuoka's theoretical work in the context of CAT, one using Kullback-Leibler information (the KL algorithm) and the other using Shannon entropy (the SHE algorithm). This paper presents an application of the optimal sequential selection method, i.e., selecting items sequentially for examinees during CAT, which is built upon a class of partially-ordered LCMs (i.e., the cognitive diagnostic models). Two new algorithms are proposed: (1) posterior-weighted KL information or PWKL method, and (2) a hybrid algorithm (HKL) which considers not only the posterior but also the distance between latent classes. Two simulation studies, one using simulated item parameters, the other with parameter estimates from real data, show that the PWKL and HKL algorithms outperformed the KL and SHE algorithms uniformly. Finally, we built the link among the algorithms by establishing equivalence between the Kullback-Leibler-information-based approaches and the Shannon-entropy-based approach, and connecting the algorithms for LCM with algorithms built upon IRT models.

**For further information:** [ycheng4@nd.edu](mailto:ycheng4@nd.edu)

## Obtaining Reliable Diagnostic Information through Constrained CAT

*Jeff Douglas, Hua-Hua Chang, and Chun Wang, University of Illinois at Champaign*

We consider how constraint weighted  $\alpha$ -stratification can be used in CAT to guarantee that sufficient diagnostic information is obtained on a set of binary latent attributes, when estimation of a unidimensional IRT ability parameter is also desired. Such applications are useful when a single score is needed, but a more fine-grained assessment of the particular skills of an examinee is also desired. Accomplishing these dual aims requires carefully constructing how a single underlying model might simultaneously contain information about a continuous latent trait and a set of binary latent attributes of a cognitive diagnosis model. Such a model is discussed and results are given illustrating how these competing models can both be thought of as valid for an exam. Implementation of constraint weighted  $\alpha$ -stratification involves identifying a priority function that combines IRT with cognitive diagnosis. Several priority functions are proposed, some based on formal measures of information, and others only utilizing knowledge of which items measure which attributes. A simulation study and results are reported, showing how utilization of information-based methods yields higher classification rates for cognitive diagnosis while achieving accurate ability estimation. Item exposure rates are also considered for all competing methods. Several new directions for future research are proposed, both for item selection and for considering when multiple latent variable models for a single dataset can be simultaneously used to extract useful information.

**For further information:** [jeffdoug@illinois.edu](mailto:jeffdoug@illinois.edu)

## Applying the DINA Model to GMAT Focus Data

*Alan Huebner, Xiang Bo Wang, and Sung Lee, ACT, Inc.*

Recent years have seen growing interest in the area of cognitive diagnostic modeling. These relatively new psychometric models seek to classify examinees as having mastered or not mastered a set of discretely defined skills, as opposed to traditional IRT models that assign examinees a continuous score measuring a broadly defined latent trait. The literature in this field contains few examples of applications of cognitive diagnostic models to real assessment data, and many of these applications use simple data sets as a means of introducing a new estimation algorithm. We attempt to fit the Deterministic Input, Noisy-And (DINA) model to assessment data for an existing test, the GMAT Focus. We discuss whether useful diagnostic information can be gleaned by applying the model to the data.

**For further information: [Alan.Huebner@act.org](mailto:Alan.Huebner@act.org)**



**G**raduate  
**M**anagement  
**A**dmission  
**C**ouncil®

*Creating Access to Graduate Business Education®*

1600 Tysons Boulevard  
Suite 1400  
McLean, VA 22102  
USA  
Phone 1-703-749-0131  
Fax 1-703-749-0169  
gmacmail@gmac.com  
www.gmac.com  
www.mba.com

Copyright © 2009 Graduate Management Admission Council® (GMAC). All rights reserved.

The Graduate Management Admission Council® is the international, nonprofit association behind the Graduate Management Admission Test® (GMAT®) used by more than 245,000 prospective MBA students and about 4,600 programs at 1,900 business schools worldwide.

Creating Access to Graduate Business Education®, GMAC®, GMAT®, Graduate Management Admission Council®, and Graduate Management Admission Test® are registered trademarks of the Graduate Management Admission Council in the United States and other countries.

GMAC® does not endorse the views of and is not affiliated with any of the speakers of this program, other than those specifically identified as representatives of the Graduate Management Admission Council®.