UNIVERSITY OF TWENTE.

IBR

RESEARCH CENTRE FOR
EXAMINATION AND CERTIFICATION

CITO now you know

International Association for
Computerized Adaptive Testing

IACAT

Advancing the Science and Practice of Human Assessment

Graduate
Management
Admission
Council®

Assessment
Systems
Corporation
www.assess.com

NWEA
Northwest Evaluation Association
Partnering to help all kids learn

MeritTrac™
India's Largest Skills Assessment Company

CollegeBoard
inspiring minds™

CITO

pi Company

RCEC
Research Center voor
Examinering en Certificering

IBR
INSTITUTE FOR
BEHAVIORAL RESEARCH

UNIVERSITY OF TWENTE.

IBR

RESEARCH CENTRE FOR
EXAMINATION AND CERTIFICATION

now you know

Hosted by the Research Centre for Examination and Certification
(RCEC: www.rcec.nl)

A partnership between Cito (www.cito.com) and

The University of Twente (www.universiteittwente.nl/en)

UNIVERSITY OF TWENTE.

IBR RESEARCH CENTRE FOR EXAMINATION AND CERTIFICATION

CITO now you know

The First International Iacat Conference Is Sponsored By:

Graduate Management Admission Council®	www.gmac.com
Assessment Systems Corporation www.assess.com	www.assess.com
CITO	www.cito.com
RCEC Research Center voor Examinering en Certificering	www.rcec.nl/en
pi Company	www.picompany.biz

UNIVERSITY OF TWENTE.

IBR RESEARCH CENTRE FOR EXAMINATION AND CERTIFICATION

CITO now you know

IBR INSTITUTE FOR BEHAVIORAL RESEARCH www.utwente.nl/ibr

NWEA™
Northwest Evaluation Association
Partnering to help all kids learn

www.nwea.org

MeritTrac™
MAGNIFYING OPPORTUNITIES
India's Largest Skills Assessment Company

www.merittrac.com

CollegeBoard
inspiring minds™

www.collegeboard.com

Graduate Management Admission Council®

Assessment Systems Corporation
www.assess.com

NWEA
Northwest Evaluation Association
Partnering to help all kids learn

MeritTrac™
India's Largest Skills Assessment Company

CollegeBoard
inspiring minds™

CITO

pi Company

RCEC Research Center voor Examinering en Certificering

IBR INSTITUTE FOR BEHAVIORAL RESEARCH

President’s Welcome

Welcome to the first conference sponsored by the International Association for Computerized Adaptive Testing (IACAT). When viewed from the perspective of the future, this conference will very likely be considered an important event in the history of CAT.

This conference is important because it is the first CAT conference to be held outside the United States. Although the first adaptive tests were developed in France over 100 years ago, the early work on the computerization of adaptive tests occurred in the United States, beginning in the late 1960s and early 1970s. The first CAT conference was held in 1975 in Washington D.C. This was followed by conferences that I hosted in Minneapolis, Minnesota in 1977, 1979, and 1982. Twenty-five years later, in 2007, Larry Rudner of the Graduate Management Admission Council asked me to co-host another conference in Minneapolis, which we repeated two years later in 2009. Although the vast majority of participants in the 1970 and 1980 conferences were from the U.S., it became clear at the 2007 and 2009 conferences that a considerable amount of CAT research and applications were taking place outside the U.S. The program for today’s conference reinforces that observation. Significant research on CAT and applications of CAT are occurring in many countries around the world.

Many people were involved in making this conference the success that I am sure it will be. Primary among those volunteers was Theo Eggen, who volunteered to host the conference. Theo has contributed many hours to conference preparations and the excellent structure of the conference is in large part due to his vision. He has been very ably assisted by Birgit Olthof, also of the University of Twente, who has worked tirelessly on the many details necessary to make conference arrangements, communicate with participants, and supervise the many details that will result in an outstanding experience for conference participants. Nate Thompson, of Assessment Systems Corporation, has contributed many hours of his time to conference preparations. Theo and Nate were assisted in structuring the program by about 15 other anonymous reviewers who reviewed and rated subsets of the paper proposals submitted. To all of these individuals a heart-felt “thank you!” Having put together a number of these conferences myself, I know how much work was involved to arrive at the outstanding schedule for this year’s CAT conference.

This conference is also important because this is the first conference organized under the auspices of IACAT. IACAT was informally founded just a year ago, at the 2009 conference, when a group of 15 or 20 attendees met to consider the formation of an international CAT organization. That meeting resulted in a small group of individuals who founded IACAT later in 2009. Foremost among that group was Cliff Donath, President of the Donath Group, who took the lead in drafting the Articles of Incorporation and the Bylaws, with major input from Nate Thompson. Once these documents were drafted, they were shared with others in the initial working group and finalized. IACAT was formally incorporated as a non-profit corporation in the State of Minnesota in January 2010. Important contributions to IACAT's early development were also made by Alan Mead, who has developed and maintained IACAT's first Web site.

IACAT's mission is summarized in our current logo as "Improving the Science and Practice of Human Assessment." Both elements of this phrase are of equal importance—we need the interaction between the scientists and those who implement CAT to move it forward as the predominant mode for measuring individual differences. We hope that you will all interact at this outstanding conference and help bridge and reinforce the connection between science and practice. If you are not a current member of IACAT, please visit our Web site, www.iacat.org, and join. If you are a member, please volunteer to help IACAT develop and move forward—we need your help and your ideas to develop future directions for IACAT. It is only with your help and assistance that we all, through IACAT and individually, will move CAT forward through conferences and other scientific and educational endeavors.

David J. Weiss, President
University of Minnesota

UNIVERSITY OF TWENTE.

IBR RESEARCH CENTRE FOR EXAMINATION AND CERTIFICATION

Cito now you know

First International IACAT Conference on Computerized Adaptive Testing

The International Association for Computerized Adaptive Testing (IACAT) is a nascent organization dedicated to advancing computerized adaptive testing (CAT) through research and education. IACAT holds its first annual conference June 7- 9, 2010. The conference takes place in Arnhem, The Netherlands. The conference, hosted by the Research Centre for Examination and Certification (RCEC: www.rcec.nl), takes place at the Conference Centre Papendal. (www.papendal.com)

Organization

The conference is organized by a committee consisting of:

- Clifford Donath, Donath Group, USA

- Theo Eggen, Cito, University of Twente, Netherlands

- Nathan Thompson, Assessment Systems Corporation,USA

- David Weiss, University of Minnesota, USA

- Dr. Natarajan Venkatesa, MeritTrac, Inc., India

- Birgit Olthof, RCEC, University of Twente, Netherlands

About IACAT

Previous CAT conferences were held in 2007 and 2009, sponsored by the Graduate Management Admissions Council. IACAT was founded at the 2009 conference, and the 2010 conference marks the first official function of the organization. Come take part in history!

To learn more about IACAT or to join, please visit www.iacat.org.

About RCEC

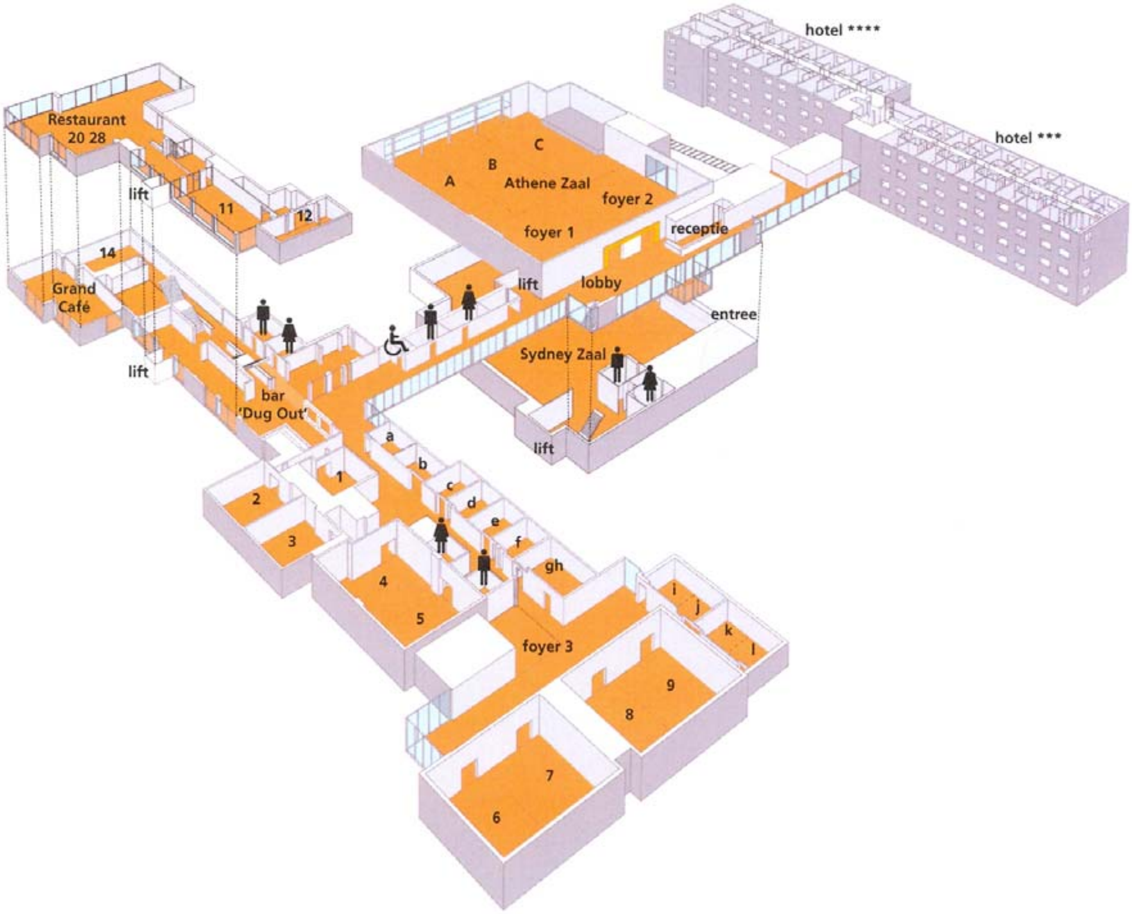
The Research Centre for Examination and Certification (RCEC: www.rcec.nl), a partnership between Cito (www.cito.com) and the University of Twente (www.universiteittwente.nl/en).

Graduate Management Admission Council
Assessment Systems Corporation www.assess.com
NWEA Northwest Evaluation Association Partnering to help all kids learn
MeritTrac India's Largest Skills Assessment Company
CollegeBoard inspiring minds
Cito
pi Company
RCEC Research Center voor Examinering en Certificering
IBR Institute for Behavioral Research

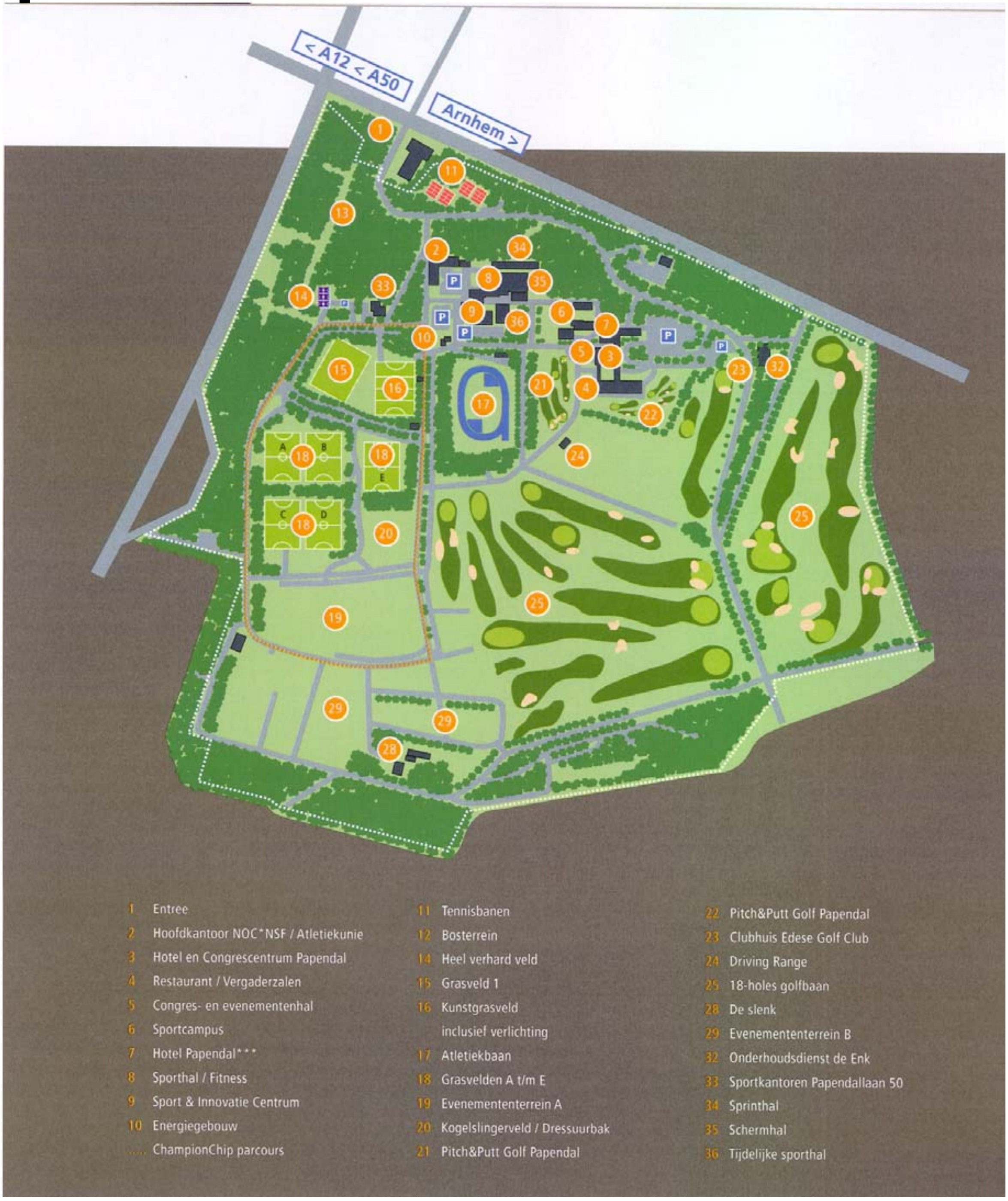
Meeting Location

Papendal Hotel and Conference Centre
Papendallaan 3
6816 VD Arnhem
The Netherlands
Phone: x31-26-483 7911
www.papendal.com

Papendal Plan inside



Papendal Plan outside


< A12 < A50
Arnhem >
P
1 Entree
2 Hoofdkantoor NOC*NSF / Atletiekunie
3 Hotel en Congrescentrum Papendal
4 Restaurant / Vergaderzalen
5 Congres- en evenementenhal
6 Sportcampus
7 Hotel Papendal***
8 Sporthal / Fitness
9 Sport & Innovatie Centrum
10 Energiegebouw
..... ChampionChip parcours
11 Tennisbanen
12 Bosterrein
14 Heel verhard veld
15 Grasveld 1
16 Kunstgrasveld inclusief verlichting
17 Atletiekbaan
18 Grasvelden A t/m E
19 Evenemententerrein A
20 Kogelslingerveld / Dressuurbak
21 Pitch&Putt Golf Papendal
22 Pitch&Putt Golf Papendal
23 Clubhuis Edese Golf Club
24 Driving Range
25 18-holes golfbaan
28 De slenk
29 Evenemententerrein B
32 Onderhoudsdienst de Enk
33 Sportkantoren Papendallaan 50
34 Sprinthal
35 Schermhal
36 Tijdelijke sporthal

Program overview

Sunday, June 6th

17.00 - 18.00	**Regfistration (foyer 3)**

Monday June 7th

08.30 - 13.00	**Registration (Foyer 3)**
09.00 - 10.15	**Preconference workshops**
Room i/j	Introduction to CAT *Nathan A. Thompson*
Room 6	Multidimensional computerized adaptive testing (MCAT) *Mark D. Reckase*
Room 7	Item selection, exposure control, and test specifications in CAT *Bernard Veldkamp*
10.15 - 10.45	**Break (Foyer 3)**
10.45 - 12.00	**Continuation of the Preconference workshops**
12.00 - 13.00	**Lunch (Restaurant 20 28)**

13.00 - 13.15	**Opening of the IACAT conference(Room 6/7)**
13.15 - 14.30	**Keynote speakers (Room 6/7)**
	The impact of computerized adaptive testing on measurement theory *Mark D. Reckase*
	How to make adaptive testing more efficient? *Wim J. van der Linden*
14.30 - 15.00	**Break (Foyer 3)**
	Paper presentations 15.00 – 16.35
Room i/j	**Contexts of CAT (concurrent session)**
15.00 – 15.20	The theoretical issues that are important to operational adaptive testing *Brian Bontempo, Steven L. Wise, Anthony R. Zara, and G. Gage Kingsbury*
15.25 – 15.45	Implementation of adaptive algorithms in CBT Systems *Mark Molenaar*
15.50 – 16.10	Adaptive testing in group level survey assessments *Andreas Oranje and Xueli Xu*
16.15 – 16.35	Enlarging an item pool by rule-based item generation. *Lolle Schakel and Annette Maij-de Meij*

Room 6	**Applications of CAT (part 1) (concurrent session)**
15.00 – 15.20	Implementing figural matrix items in computerised adaptive testing system *Tay Poh Hua, Raymond Fong, Low Sik Kuan and Chee Meng On*
15.25 – 15.45	Short forms versus CAT for the assessment of patient reported outcomes measures *Richard C. Gershon, David Cella and Seung Choi*
15.50 – 16.10	Validity of CAT in personnel selection *Sara Gutierrez, Darrin Grelle & Mike Fetzer*
16.15 – 16.35	Adaptive rule-based intelligence testing *Jonas Bertling & Heinz Holling*
Room 7	**Issues in ability estimation (concurrent session)**
15.00 – 15.20	Competence´s initial estimation in computer adaptive testing *Félix Castro, Joel Suárez and Raul Chirinos*
15.25 – 15.45	Guess again: The effect of correct guesses on scores in an operational CAT program *Eileen Talento-Miller, Kyung T. Han, Fanmin Guo*
15.50 – 16.10	Ability estimation and test ending strategies of CAT for achievement testing on different samples *Ilker Kalender and Giray Berberoglu*
16.15 – 16.35	Effect of θ estimation method and starting value on the recovery of θ *Rick Guyer and David J. Weiss*

Paper presentations 16.50 – 18.00

Room i/j **Item selection (Part 1) (concurrent session)**

16.50 – 17.10 CAT item selection and person fit: Predictive efficiency and detection of atypical symptom profiles
Barth B. Riley, Michael L. Dennis, and Kendon J. Conrad

17.15 – 17.35 *Adaptive tests of adjustable difficulty*
Huub Verstralen

17.40 – 18.00 Effects of computerized adaptive testing and mean response probability on test-taking motivation and performance
Regine Asseburg

Room 6 **Utilizing time data (concurrent session)**

16.50 – 17.10 Using response times to improve measurementprecision for examinees with aberrant response behaviors in computerized adaptive testing
Jyun-Hong Chen, Shu-Ying Chen, Chuan-Ju Lin

17.15 – 17.35 A smple method for controlling test time intensity in CAT
Fanmin Guo

17.40 – 18.00 When does item duration become a relevant variable in measurement?
Dr. Annette Maij-de Meij and MSc. Lolle Schakel

Room 7 **New methods (concurrent session)**

16.50 – 17.10 Using the Rasch model-based LLTM for composing item difficulties on demand – the test Alpha-Numerical TOPologies
Klaus D. Kubinger and Nina Heuberger

17.15 – 17.35 A Kalman filtering approach to computerized adaptive testing
P.W. van Rijn

17.40 – 18.00 Bayesian optimal design for the Rasch model and linear logistic test model
Heinz Holling

19:00 **Dinner (Restaurant 20 28)**

Tuesday, June 8th

09.00 - 10.15 **Keynote speakers (room 6/7)**

Clinical and medical applications of computer adaptive testing
Otto B. Walter

Computer adaptive testing for small scale programs and instructional systems
Lawrence M. Rudner

10.15 - 10.45 **Break (Foyer 3)**

10.45 - 12.00 **Poster Presentations (room 6/7)**

Does the theory work in practice?
Hilary Ferral

CML estimations in multistage testing
Robert Zwitser and Gunter Maris

Assessing the critical thinking ability of undergraduate students in a Nigerian University
Dr. H. O. Owolabi, Mrs. C. O. Owolabi, Dr. Bisi Onasanya and Dr. Mrs. R. O. Oduwaiye

An adaptive scheme for the dynamic rasch calibration of pilot items
Rense Lange

Teate depression inventory: A rasch derived self-report inventory of depression
Balsamo, Michela, Sergi, Maria Rita, Saggino, Aristide and Giuseppe Giampaglia

Computer generated item analysis based on three-parameter Logistic model as applied to chemistry achievement test
Susan C. Unde and Chita P. Evardone

How does the equal step size affect the number of items admistered in computerized adaptive testing
Atilla Halil ELHAN and Derya ÖZTUNA

Teacher and student experiences with the use of an adaptive test (WISCAT-pabo)
Mia van Boxel and Jeanine Treep

Designing a computerized adaptive test to measure verbal reasoning
Gabriela Lozzia, María Ester Aguerri, Facundo Abal and Horacio Attorresi

Selection of common items in full metric calibration for the development of CAT item banks
Jieun Lee and David J. Weiss

A comparison of different parameter estimation methods
Isabel Cañadas, Sonia Tirado and Rebeca Bautista

Omitted responses: The effect on item parameters
Sonia Tirado, Rebeca Bautista and Isabel Cañadas

The sample size effect on item parameters
Rebeca Bautista, Isabel Cañadas and Sonia Tirado

Effects of medium on writing assessment of learners of chinese as a foreign language
Yu Zhu (Ph.D.) and Andy SL Fung (Ph.D.)

Estimating item response theory based item parameters when response matrix is quite sparse
Vipin Chilana

Development of a physical function CAT
Man Hung, David J.Weiss and Charles L. Saltzman

Constructing a new fluid intelligence test: An IRT approach
Romanelli R., Saggino A. and Weiss D. J.

12.00 - 13.00 **Lunch (Restaurant 20 28)**

Paper presentations 13.30 – 15.05

Room i/j **Multidimensional CAT (concurrent session)**

13.30 – 13.50 A general multidimensional computerized adaptive testing platform in health outcomes measurement
Seung W. Choi, Richard C. Gershon, and Dave Cella

13.55 – 14.15 Benefits of multidimensional adaptive testing from a practical point of view
Ulf Kröhne and Ivailo Partchev

14.20 – 14.40 The bifactor model and its application to multidimensional computerized Adaptive Testing
Dong Gi Seo and David J. Weiss

14.45 – 15.05 Multiple-category classification with multidimensional adaptive testing in large-scale assessments
Nicki-Nils Seitz and Andreas Frey

Room 6 **CAT for Classification (concurrent session)**

13.30 – 13.50 Adaptive sequential mastery testing using the Rasch model and Bayesian sequential decision theory.
Hans J. Vos and Cees A.W. Glas

13.55 – 14.15 Effects of different termination criteria on classification consistency in CAT
Nam Keol Kim

14.20 – 14.40 Using the sequential probability ratio test when items and respondents are mismatched.
Maaike van Groen & Angela Verschoor

14.45 – 15.05 Nominal error rates in computerized classification testing
Nathan A. Thompson

Room 7	**CAT in Education (concurrent session)**
13.30 – 13.50	Orthography testing and CAT - on the way to establish individual orthography learning *Sarah Frahm*
13.55 – 14.15	Adapt your listening test to your level *Klaas Schreuder, Angela Verschoor, and Niels Veldhuijzen.*
14.20 – 14.40	WISCAT-pabo: Computerized adaptive testing of arithmetic skills of first-year students at primary school teacher training colleges *Gerard J.J.M. Straetmans and Theo J.H.M. Eggen*
14.45 – 15.05	DESIGNING & Using adaptive tests for large scale formative assessment *John Winkley*
	Paper presentations 15.20 – 16.30
Room i/j	**Item selection (Part 2) (concurrent session)**
15.20 – 15.40	Item selection criteria in CAT based on predictive distributions *Mathilde Huisman-van Dijk and Frans Kamphuis*
15.45 – 16.05	Content control with the maximum priority index method in multidimensional adaptive testing *Andreas Frey and Ying Cheng and Nicki-Nils Seitz*
16.10 – 16.30	The maximum priority index for generalized partial credit model in computerized adaptive testing *Zeng Lingyan*

Room 6	**Conformation testing and assessment of change (concurrent session)**
15.20 – 15.40	Unproctored internet test verification: Using adaptive confirmation testing *Guido Makransky and Cees Glas*
15.45 – 16.05	Accepting the null: Determining no change within the adaptive measurement of change *Steven W. Nydick and David J. Weiss*
16.10 – 16.30	Measuring individual growth curves with computerized adaptive testing *Shannon M. Von Minden and David J. Weiss*
Room 7	**Collatoral information in CAT (concurrent session)**
15.20 – 15.40	A Bayesian approach for introducing empirical information in CAT *Mariagiulia Matteucci*
15.45 – 16.05	*Text classification in prior selection of CAT* *Qiwei He*
16.10 – 16.30	*Robust item selection in computerized adaptive testing* *Bernard Veldkamp*
17:00	**Social activity Sponsored by Cito** **(Pitch & Putt at number 22 on map)**
19:00	**Dinner (Restaurant 20 28)**

Wednesday, June 9th

09.00 - 10.15 **Keynote speakers (Room 6/7)**

Bootstrapping an item bank
Cees. A. W. Glas and Guido Makransky

Computerized classification testing to predict an observable outcome
Matthew Finkelman, Yulei He, Giles Hooker, Wonsuk Kim, Robert Keller and Barbara Gandek

10.15 - 10.45 **Break (Foyer 3)**

Paper presentations 10.45 – 12.20

Room i/j **Item bank development (concurrent session)**

10.45 – 11.05 Pushing the limits of an adaptive test: Interaction between the test population and the item pool
Steven L. Wise, Brian Bontempo, and G. Gage Kingsbury

11.10 – 11.30 Optimal calibration designs for computerized adaptive testing
Angela Verschoor

11.35 – 11.55 Flexibly and accurately evaluating item pool adequacy and optimization for linear-to-CAT conversion using optimization
Xiang Bo Wang and Xinhui Xiong

Room 6/7 **Applications of CAT (part 2) (concurrent session)**

10.45 – 11.05 A comparison of item exposure, test information, and theta recovery for LOFT, CAT, and fixed form tests.
Kirk A. Becker, Brian Bontempo, Phil Dickison, and James S. Masters

11.10 – 11.30 Opportunities on adaptive testing in India
Dr. V. Natarajan and Mr. Rajeev Mennon

11.35 – 11.55 A dynamic way of equating for CAT practice
Zhang Quan

Room k/1 **Item Exposure (concurrent session)**

11.45 – 11.05 Conditional multiple maximum exposure rates in CATs.
Juan Ramon Barrada, Julio Olea, and Francisco J. Abad

11.10 – 11.30 The on-line procedure for simultaneous control of item exposure and test overlap in variable length computerized adaptive testing
Chia-Ling Hsu and Wen-Chung Wang

11.35 – 11.55 Controlling on-line item exposure and test overlap in CAT with the ability-based guessing model
Shiu-Lien Wu, Wen-Chung Wang, and Shu-Ying Chen

Room 6/7

12.30 - 12.45 **Closing of the IACAT conference**

12.45 - 13.30 **Lunch (Restaurant 20 28)**

Program details

Monday, June 7th

08.30 - 09.00 **Registration**

09.00 – 12.00 **Preconference workshops**

Introduction to CAT

Nathan A. Thompson

This workshop will provide a conceptual introduction of CAT and the item response theory that supports it. It will begin by describing the item response function and how it is used to evaluate item information and provide estimates of examinee ability. It will then discuss the five components necessary to produce a CAT.

Multidimensional Computerized Adaptive Testing (MCAT)

Mark D. Reckase

This workshop provides practical information about how to develop and implement a multidimensional item response theory-based computerized adaptive testing system (MCAT). The workshop will cover the following topics:

1 Defining the multidimensional space for the MCAT.
2 Developing an item pool for the MCAT including linking calibrations to achieve item pools of the necessary size for implementation of a MCAT procedure.
3 Alternative MCAT methodologies.
4 Issues related to design and implementation of the procedures.

The workshop will be based on the last few chapters of the book Reckase, M. D. (2009). *Multidimensional Item Response Theory*. New York: Springer. Also, MATLAB programs that have been developed for simulating MCAT procedures will be shared with the workshop participants.

Item Selection, Exposure Control, and Test Specifications in CAT

Bernard Veldkamp

The process of Computerized Adaptive Testing (CAT) has five basic steps: (1) an initial ability estimate is made for the candidate, (2) an item is selected, (3) the item is administered, (4) the ability estimate is updated, and (5) steps 2 - 4 are repeated until a stopping criterion has been met. Although seemingly straightforward, there are a number of issues that have to be dealt with for each of these steps. In this workshop we will focus on step (2) of the algorithm, the selection of the next item in computerized adaptive testing. We will deal with three important issues:

1. Which item selection criterion to apply.
2. How to deal with exposure control.
3. How to deal with test specifications.

Several item selection rules have been proposed in the literature to deal with the first issue. In the workshop, the focus will be on Maximum Fisher information, Fisher interval information, Kullback-Leibler information and several Bayesian item selection criteria. How do they differ, what are the advantages and disadvantages, and which of them should be applied?

The Sympson Hetter method is most commonly applied for exposure control. This method will be introduced and several modifications of the method will be discussed. Alternatives such as the Alpha-stratified method, the Progressive method, and the Item eligibility method will also be compared. Finally, several approaches for implementing test specifications in the item selection algorithm will be presented to address the third issue. This workshop will be a mix of theory, discussion, sharing experiences, and exercises. It will deal with some more advanced issues in computerized adaptive testing. In order to participate, you need to have some basic knowledge of Item Response Theory models, such as the Rasch-model, the 2PLM, the 3PLM, and polytomous IRT models as well as an understanding of the basic principles underlying CAT.

12.00 - 13.00	**Lunch and registration**
13.00 - 13.15	**Opening**
13.15 - 14.30	**Keynote speakers**

The impact of computerized adaptive testing on measurement theory

Mark D. Reckase

Computerized adaptive testing (CAT) is a well known methodology among educational and psychological measurement professionals. However, the impact of the concepts behind CAT on the broader fields of educational and psychological measurement is underappreciated. An argument will be made in this presentation that the work on the development of practical CAT procedures was at least partially responsible for a paradigm shift in the theory of practice of educational and psychological measurement. That paradigm shift will be described and examples of the influence of the paradigm shift will be presented.

How to make adaptive testing more efficient?

Wim J. van der Linden

As a rule of thumb, adaptive testing requires some 50% of the test length to achieve the same efficiency of ability estimation as a linear test. In this presentation, I will use hierarchical IRT modeling to show that further increase of efficiency is quite possible. In one application, a hierarchical framework with an IRT model for the responses and an additional model for the response times on the test items is used to improve ability estimation and item selection in the adaptive test. In another application, a hierarchical IRT approach is used to sequence a test battery, allowing the use of information from an earlier subtest to adaptively pick the next subtest. Both applications are illustrated for an adaptive version of the Law School Admission Test (LSAT).

14.30 - 15.00 **Break**

15.00 - 16.35 **Paper Presentations**

Contexts of CAT

The theoretical issues that are important to operational adaptive testing

Brian Bontempo, Steven L. Wise, Anthony R. Zara, and G. Gage Kingsbury

This set of coordinated presentations will investigate the theoretical questions regarding adaptive testing that affect operational adaptive tests most. These include

1 What do we care about related to exposure control?
2 How should we really select items to help test users?
3 How should we field test items?
4 How do we look at growth across time with a series of adaptive tests?
5 How do item pools perform if the testing population is changing?
6 What happens when the test breaks?
7 How do we examine the validity of individual test performance?

Most of these questions have been addressed in the academic literature, but not with an eye to application. This session asks whether and to what extent operational practice can help inform the next generation of research that needs to be done. This session includes folks who have been moving from theory to practice and answering each of the questions above in a variety of innovative ways. Presenters bring experience that includes the development of high and low-stakes adaptive tests used in education, certification, and licensure. This session will discuss the approaches that have been taken in a variety of settings, and include a discussion of the pros and cons of different practical solutions. The session will also identify theoretical questions that really need to be answered quickly to allow the continued rigorous development of adaptive tests that achieve the goal of equiprecise measurement in operational context.

Implementation of adaptive algorithms in CBT systems

Mark Molenaar

The level of complexity of CBT's is increasing rapidly in multiple dimensions. Not only in the field of CAT, but also in the field of application- and multimedia-development. New adaptive algorithms, next generation test-delivery engines and rich multimedia content need to be "mashed-up" into one psychometrically sound, maintainable, durable and visually attractive CBT. This requires the skill and collaboration of multiple distinct disciplines. The key to multidisciplinary development is a modular, extensible architecture and clear (technical) interfaces. This paper will address the specifics of a simple technical interface for an adaptive "plug-in" and illustrate how to extend its application.

Adaptive testing in group level survey assessments

Andreas Oranje and Xueli Xu

In this presentation we will discuss a practical framework for the application of adaptive testing to educational survey assessments, followed by an implementation of a multi-stage test (MST) in the National Assessment of Educational Progress (NAEP). Survey assessments such as TIMSS, PISA, and NAEP are generally concerned with measurement accuracy at the group level. Subsequently, the potential benefits of adaptive testing are quite different from traditional applications and may include simplification of estimation methodologies (particularly in relation to population models), reduction of secondary biases, accommodation of students with disabilities or language deficiencies, and improvement of engagement in low-stakes assessments.

Enlarging an item pool by rule-based item generation

Lolle Schakel and Annette Maij-de Meij

A CAT needs a large item pool. To enhance item writing efficiently, insight into the relation between item characteristics and item parameters is required. One can gain in efficiency by effectively writing good quality items with the aimed difficulty level, for example by rule-based item generation (automatically or manually). A model-based approach for a Series of Numbers test, using regression models, is followed. The model is used to predict item parameters based on information with respect to various item characteristics. Furthermore, the accuracy of item parameter estimation by item writing experts is evaluated.

Applications of CAT (part 1)

Implementing figural matrix items in computerised adaptive testing system

Tay Poh Hua, Raymond Fong, Low Sik Kuan, and Chee Meng On

In this study, an item bank of figural matrix items, similar to Raven's Standard Progressive Matrices (SPM) was created. Two CAT Prototypes (one starts with an easy item, while the other starts with an average item), and a Computer Based Test (CBT) Prototype were generated and administered via the FastTEST to three groups of Primary 2 (equivalent to Grade 2 pupils who are about 8 years in age) pupils in our country. This study was designed to ascertain the comparability of the prototypes and the importance of the progressive nature of SPM in the measurement of pupils' fluid ability.

Short forms versus CAT for the assessment of patient reported outcomes measures

Richard C. Gershon, David Cella, and Seung Choi

The U.S. National Institutes of Health (NIH) Roadmap Initiative: Patient Reported Outcomes Measurement Information Systems (PROMIS) has created IRT calibrated item banks, computerized adaptive tests and short forms to assess a wide array symptoms and functional domains associated with physical, mental and social health. Item banks have been created through using a rigorous methodology including literature reviews, patient interviews and numerous qualitative item review processes. Calibration and validation activities have to date encompassed over 20,000 subjects in both general population and disease-specific samples. This presentation will report on the results of this work to date with a particular emphasis on the strengths and weaknesses of utilizing CAT versus various short form versions of each measure. The discussion will include issues of time and burden and accuracy for each instrument type across the range of what is being measured.

Validity of CAT in personnel selection

Sara Gutierrez, Darrin Grelle & Mike Fetzer

This presentation will review the development of CAT-based cognitive ability and personality assessments and provide examples of the benefits of utilizing computer adaptive assessments in a selection context. In addition, local and meta-analytic evidence indicating strong correlations between test scores and job performance criteria across multiple samples and industries will be presented. Finally, a unique method of leveraging computer adaptive testing methods to further mitigate the risks of unsupervised testing will be described. The use of CAT for personnel selection is showing great promise. Researchers and practitioners will benefit from this discussion of a new method of pre-employment testing.

Adaptive rule-based intelligence testing

Jonas Bertling & Heinz Holling

Adaptive tests for general intelligence will be presented. Rules for three types of items, Latin Squares, Figural Analogies have been derived and allow for constructing these items on the fly. The items are Rasch scalabe and difficulties of the rules were calibrated by linear logistic test models. Person parameters are determined using the WARM estimator.

Issues in ability estimation

Competence´s initial estimation in computer adaptive testing

Félix Castro, Joel Suárez and Raul Chirinos

In the experiments we apply the Fuzzy Inductive Reasoning (FIR) methodology to perform initial examinee's ability estimation in CAT. Using FIR we also perform a feature relevance determination with the goal to identify the features that most influences the examinee's ability estimation. Furthermore, we apply a novel rule-extraction algorithm based on fuzzy logic (LR-FIR) to extract examinee's behaviour patterns and express them in an actionable, and easy interpretable way. The obtained results in applying the FIR and LR-FIR algorithms, are very successful and encouraging because offer valuable information about the capacity of these tools in the CAT's estimation ability process.

Guess again: The effect of correct guesses on scores in an operational CAT program

Eileen Talento-Miller, Kyung T. Han, Fanmin Guo

Previous theoretical research shows that recovery of ability estimate on computerized adaptive tests after initial correct responses depends on factors such as the item selection method and test length. Data from an operational testing program and actual examinees were used to simulate the effect of strings of correct responses in different positions on the exam. Differences interacted relative to string length, position, section, and examinee ability. Regardless of string position, average score differences were less than a standard error. Findings refuted the myth that the beginning items are the most important.

Ability estimation and test ending strategies of cat for achievement testing on different samples

Ilker Kalender and Giray Berberoglu

The purpose of the present study is to compare two different ability estimation procedures (Maximum Likelihood and Bayesian EAP methods) and two test ending criteria (fixed item number and threshold of standard error) of CAT using real data. To detail the analyses, follow-up comparisons were also conducted on different student groups. These cognitively varying groups show different test and personal characteristics such as missing response rates, selective answering, computer familiarity, etc. /Findings pointed out that different ability estimation methods and test ending criteria are needed for different subgroups. Sources of needs for different CAT strategies for subgroups will be discussed.

Effect of θ estimation method and starting value on the recovery of θ

Rick Guyer and David J. Weiss

This study focused on how θ estimation method affects the recovery of θ for a computerized adaptive test (CAT). Maximum likelihood estimation (MLE) cannot obtain finite θ estimates when the response pattern is not mixed. This study examined three different methods for handling this problem in CAT. The first method is to adjust the θ estimate by an arbitrary value (e.g., 1) while the response pattern is not mixed. Alternatively, weighted maximum likelihood (WML) or Bayesian scoring (expected a posteriori; EAP) can be used for scoring while the response pattern is not mixed. MLE scoring was used for the remainder of the CAT once a mixed response pattern was obtained. To serve as a baseline, WML and EAP scoring were also used for the duration of the CAT. The starting θ value was also an independent variable in this study. The starting θ value was varied from -2 to +2 in one unit increments. This study used a monte carlo design with generating θ values of -2, -1, 0, +1, and +2. There were 1000 replications performed per cell. Recovery of θ was indexed by bias, standard error, and root-mean-square error at CAT lengths of 15, 25, 35, and 50 items. Item parameters were generated according to the following distributions: a ~ log-normal(–0.223, 0.2), b ~ U[–3.5, 3.5], c ~ N(.20, .02). The mean of a in the logistic metric was about 0.82 with a standard deviation of 0.15.

Previous research that examined the effect of θ estimation method on the recovery of θ in CAT did not examine the effect of varying the starting θ value. In addition, since WML can obtain finite θ values for non-mixed response patterns it was examined as an alternative to EAP early in the CAT. No previous research has examined the difference between WML and EAP as estimation methods in CAT when the response pattern is not mixed.

16.35 – 16.50 Break

16.50 – 18.00 Paper Presentations

Item selection (Part 1)

CAT item selection and person fit: Predictive efficiency and detection of atypical symptom profiles

Barth B. Riley, Michael L. Dennis and Kendon J. Conrad

This paper focuses on the predictive efficiency and sensitivity of person fit statistics to detect atypical response patterns in computerized adaptive testing (CAT) as a function of number of items administered and item selection method. We compared maximum Fisher's information item selection with an approach loosely based on Linacre's (1995) Bayesian "maximum falsification" method. Data from 4,360 individuals presenting to substance abuse treatment who completed a measure of internal mental distress were used to perform a series of CAT simulations. Results generally suggested that the modified Bayesian procedure resulted in improved sensitivity and predictive efficiency of person fit indices and improved ability to detect persons with atypical symptom profiles.

Adaptive tests of adjustable difficulty

Huub Verstralen

An idea originally put forward by Eggen and Verschoor (2006) has been further developed by replacing the information function by a selection function with adjustable selectivity. By introducing higher selectivity than the information function the risk of suboptimal item choice could be reduced, and the wanted difficulty could be better approached. Additionally, the analysis of bias of the ability estimate was redone with estimated parameters. It was found that higher selectivity and large calibration samples (>4500) help to avoid severe bias and to approach the results of Eggen and Verschoor (o.c.) with true parameter values.

Effects of computerized adaptive testing and mean response probability on test-taking motivation and performance

Regine Asseburg

Because of its high measurement efficiency, computerized adaptive testing (CAT) is a promising test algorithm to be implemented in large-scale assessments. However, the effects of CAT on test-taking motivation, which is positively correlated with performance in low-stakes test situations, are not yet thoroughly understood. Based on expectancy-value theory of motivation, this 2x2 experiment investigates the effects of test algorithm (CAT/fixed item testing) and mean response probability (medium/low) on test-taking motivation and mathematical performance, assuming differential effects for participants with high/low mathematical ability (N = 343 ninth graders). Results and implications for the application of CAT in educational testing are discussed.

Utilizing time data

Using response times to improve measurement precision for examinees with aberrant response behaviors in computerized adaptive testing

Jyun-Hong Chen, Shu-Ying Chen, Chuan-Ju Lin

The efficiency of CAT would not be held when examinees had aberrant response behavior. Aberrant response behavior means that examinees answered the items based on item pre-knowledge or rapidly guessing, rather than on their ability. The purpose of this study is to improve measurement precision for examinees with aberrant response behavior, where the aberrant responses were excluded based on a procedure proposed by van der Linden & van Krimpen-Stoop (2003) with consideration of response times. Results indicated that the precision of ability estimates were significantly improved when response times were taken into account in the detection of aberrant responses.

A simple method for controlling test time intensity in CAT

Fanmin Guo, Director of Psychometric Research, Graduate Management Admission Council, USA

In computerized adaptive tests (CAT), examinees receive tailored tests. Some tests built by computers might have more time-intensive items, thus taking longer to finish than other tests. The consequences might disadvantage some examinees unfairly and allow others unfair benefits. That is especially the case in fixed-length and fixed-time CAT programs.

In this paper, the author describes a simple compensatory method for controlling time intensity in CAT tests, discusses the estimation of item time intensity parameters, and reports a simulation study of applying this method to a CAT program in comparison with a baseline simulation with no test time controls.

When does item duration become a relevant variable in measurement?

Annette Maij-de Meij and Lolle Schakel

How can test item duration and so total test time in a CAT be reduced responsibly? For an adaptive test for cognitive ability the question posed itself whether time limits could be further reduced responsibly, without having to adjust or recalibrate the item pool, and how that should be investigated. Empirical data with artificial time reduction have been studied for the influence of time reduction on p-values and item parameter estimates. Results will be presented, (dis)advantages of the applied method as well as directions for future research will be discussed.

New methods

Using the Rasch model-based LLTM for composing item difficulties on demand – the test Alpha-Numerical TOPologies

Klaus D. Kubinger & Nina Heuberger

In example of the reasoning test ANTOP (*Alpha-Numerical TOPologies*; Kubinger & Heuberger, in prep.) some revivaled ideas are given of how to compose an item's Rasch model item difficulties by applying the LLTM. That is, item difficulty parameters are decomposed by a weighted sum of so-called operation parameters, these representing the difficulty of some hypothesized item solution components. If the data's likelihood based on the LLTM and consequently on just a "handful" of operation parameters comes close to the data's likelihood based on the Rasch model and as many item difficulty parameters as there are items of the test then the hypothesized solution components were verified (cf. e.g. Kubinger, 2008, 2009).

A Kalman filtering approach to computerized adaptive testing

P.W. van Rijn

In this presentation it is discussed how Kalman filtering techniques can be used in the setting of computerized adaptive testing (CAT). It is illustrated how an adaptive testing system can be built from scratch, that is, without knowledge of item and person parameters. For the illustrations, a generalized Kalman filtering technique is used as described in Fahrmeir and Wagenpfeil (1997). It is demonstrated how the estimation of item parameters can be optimized for the 2-PL model in the build-up phase of a computerized adaptive test. Particularly, adaptive estimation of discrimination parameters is addressed, because their estimation can be troublesome (Hambleton, Swaminathan, & Rogers, 1991).

Bayesian optimal design for the Rasch model and linear logistic test model

Heinz Holling

In this contribution optimal designs for the Rasch model and linear logistic test model will be presented based on a Bayesian approach. First, both models are embedded in a particular generalized linear model. Then, locally optimal designs can be constructed with respect to item calibration. Since these designs are quite restrictive the Bayesian framework is applied. Optimal designs using different prior distributions, e. g. normal and logistic distributions, will be derived.

19.00 Dinner

Tuesday, June 8th

09.00 - 10.15 Keynote speakers

Clinical and medical applications of computer adaptive testing

Otto B. Walter

The majority of theoretical and practical contributions concerning the application of computer adaptive tests (CATs) has been focused on the ability and achievement testing context. Many of the advantages of CATs seem to be well suited for the assessment of clinical constructs or health-related measures but the era of CATs in health-related or clinical measurement has just begun. The talk highlights recent developments and discusses practical issues such as computer adaptive testing on mobile devices.

Computer adaptive testing for small scale programs and instructional systems

Lawrence M. Rudner

Based in item response theory, much of the research and development for computer adaptive testing has been has been restricted to programs with large item banks and large numbers of examinees. This paper will present approaches to CAT that lift these requirements. These measurement decision theory approaches are simple to implement and are well suited for small certification programs and as embedded tools in intelligent instructional systems. Publically available software for simulating data, calibrating questions, and scoring examinees will also be presented.

10.15 - 10.45 Break

10.45 - 12.00 Poster presentations

Does the theory work in practice?

Hilary Ferral

Recently, a new computer adaptive assessment tool has been developed in New Zealand to assess adults with lower levels of literacy and numeracy skills.

We present the results of a set of simulations designed to evaluate the tool's combined power of a large item bank and item selection algorithms to provide appropriate CAT assessments for test-takers of varying ability. While aggregate statistics regarding the performance of the algorithms are interesting and provide critical information, we present here some graphics which help to visualise the processes at play, gain insight into a test-taker’s experience, and assess how well the assessment tool will work in practice.

Inverted computer-adaptive rasch measurement: Prospects for virtual and actual reality

Matt Barney

Computer-adaptive testing is significantly shorter and more secure than alternative approaches because it tailors the set of items administered to each unique person. But CATs can still be tedious and items misappropriated. Further, it is possible for people to misrepresent themselves in self-report measures. A new approach called “Inverted Computer Adaptive Testing” based on Rasch Measurement, is proposed for Virtual Reality environments that may hold promise for overcoming some of these challenges in an engaging and realistic environment that mimics real life. What adapts in ICAT is the certainty of a person’s location on a dimension, based on actions in a computer-generated world. In ICAT behaviors elicited become items whose locations dynamically update confidence intervals for person location parameters. Such an approach has potential to be used to give dynamic feedback that may improve both measurement and development effectiveness. Even though Virtual Reality can be costly, potential benefits for high-stakes environments include 1) more engaging experiences; 2) reducing item over-exposure; 3) adjusting for bias; 4) improving feedback to enhance transfer; 5) enabling new unobtrusive item types including behavioral residuals; 6) reducing misrepresentation; and 7) providing new electronic methods for anthropological / observational studies. A Monte Carlo study is presented to illustrate the potential of the approach to better detect measurement distortions by combining fit statistics with Taguchi process capability estimates from Industrial/Systems Quality Engineering.

CML estimations in multistage testing

Robert Zwitser and Gunter Maris

A multistage test is probably the simplest adaptive test possible, where students are administered different blocks of items depending on the responses to earlier blocks of items. It has often been observed that for calibration purposes the method of CML is not applicable with data from a multistage test. Instead, the method of MML is mostly used.However, it will be shown that the MML estimates are significantly biased if the population distribution is not correctly specified. As a solution, it will be shown that, using the correct conditional likelihood, the method of CML provides unbiased estimates. The results will be illustrated with simulated data.

Assessing the critical thinking ability of undergraduate students in a Nigerian University

Dr. H. O. Owolabi, Mrs. C. O. Owolabi, Dr. Bisi Onasanya and Dr. Mrs. R. O. Oduwaiye

The development of critical thinking ability has become a common theme of national education objectives. It is a skill for successfully living. There are two perspectives to its development: subject matter approach integrates teaching-learning and testing with instruction while direct approach teaches the skill and then uses independent tests. Nigerian education seems to have adopted the subject matter approach. Instruments for direct test of critical thinking are not commonly used and this raises a demand. A computerized adaptive critical thinking test was developed and validated for use among Nigerian undergraduate students. The instrument was found to be valid and reliable.

An adaptive scheme for the dynamic rasch calibration of pilot items

Rense Lange

To efficiently build large item pools, new “pilot” items (i.e., items with unknown parameters) can be calibrated dynamically during CAT using “known” (i.e., previously calibrated) items. During simulated Rasch CAT using 250 “known” and 1000 “pilot” items, every fifth administered was a pilot item. Pilot items were selected either at random, or using maximum information based on dynamically updated PROX estimates. As expected, the maximum information approach outperformed random selection. Surprisingly, the final dynamic PROX estimates were superior to UCON estimates over all available data (i.e., known and pilot items). Thus, scaling sparse data may pose particular challenges.

Computer generated item analysis based on three-parameter logistic model as applied to chemistry achievement test

Susan C. Unde and Chita P. Evardone

The annual Chemistry Achievement Test (CAT) is a division-wide test conducted among the public high schools in Iligan City to evaluate the performance of students in Chemistry and likewise used as an instrument to measure teaching efficiency.

This study was conducted to evaluate the CAT with the objective of obtaining a pool of valid good questions in Chemistry. A computer generated item analysis using the Item Response Theory was performed based on the Three-parameter Logistic Model. Findings revealed that 25% of the items of the CAT conducted are not valid and thus the test was considered unreliable on this basis. A calibrated CAT was recommended.

How does the equal step size affect the number of items admistered in computerized adaptive testing

Atilla Halil ELHAN and Derya ÖZTUNA

The aim of this study was to explore the effect of the width of adjacent thresholds (step size) on the number of items administered in CAT application. We have simulated four different item banks each contains 26 items with five responses according to Rasch model and arranged difficulty parameters as Uniform (-5,5). We have changed the width of step size to 0.5, 1, 2, and 4 for the item banks. Then, the WinGen2 was used to simulate the responses of 1.000 simulees distributed as Uniform (-5,5). The CAT assessments reduced the median number of items administered from 26 to 5, 6, 9 and 16 for the item banks, respectively.

Teacher and student experiences with the use of an adaptive test (WISCAT-pabo)

Mia van Boxel and Jeanine Treep

In this poster presentation the following topics will be discussed:

- What do teachers and students like best: adaptive or linear tests? What is the reason for this preference?
- Which formats can be used best for reporting the results of an adaptive test?
- Which measures must be taken to make the implementation of an adaptive test acceptable and successful? Subjects here are for instance fixed length and conversion of “ability estimates” to “marks or grades”.
- What about the right of students to review the test, if the test is only manifest at the moment of examination because there is no test, there is only an item bank.

During the poster session there will be a demo of the WISCAT-pabo, so the vistitors can try it and experience adaptivity by themselves .

Teate depression inventory: A rasch derived self-report inventory of depression

Balsamo, Michela, Sergi, Maria Rita, Saggino, Aristide and Giuseppe Giampaglia

The Teate Depression Inventory (TDI; Balsamo, 2006) was a 21-item unidimensional self-report scale of depression, constructed using a modified version of the Rasch one-parameter model (Rasch, 1960-1980), proposed by Andrich (1988). The Rasch model was elaborated by values matrix produced by 529 subjects (either clinical subjects or healthy subjects) and 51 items, elicited on the basis of the diagnostic criteria for Major Depression Episode of the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV-TR; APA, 2000). Following Rasch analysis, we selected 21 items by removing items on the basis of three criteria: model fit, individual item fit, and item residuals (Giampaglia, 2008).
It has some advantages, as ordered sequence of the thresholds of each of the 21 items; computation of a total score by summing the categories response to the items; good discriminative ability between normal and clinical subjects. Thus, it could be a valid screening instrument in that it sensitively discriminates between outpatients and normals.

Designing a computerized adaptive test to measure verbal reasoning

Gabriela Lozzia, María Ester Aguerri, Facundo Abal and Horacio Attorresi

This work displays the current stage of progress in developing an item bank for designing a Computerized Adaptive Test to measure Verbal Reasoning. The general features of its items regarding to their contents and IRT-parameters, and the criterions to be followed for designing a CAT are shown. Some interesting achievements are expected to come from this CAT, including providing a useful instrument for the educative field, introducing a new methodology in our country -where IRT is still only moderately developed and applied- and the possibility of exploring the reaction of university students when they are taking this kind of test.

Selection of common items in full metric calibration for the development of CAT item banks

Jieun Lee and David J. Weiss

A CAT item bank typically requires the administration of subsets of different items to different groups of examinees. Full-metric concurrent calibration is a method that can be used to link the IRT item parameters onto a scale that reflects the full □ range for the total group of examinees. This simulation study examined the performance of full-metric concurrent calibration focusing on the anchor test design under the 3-parameter model. Number of groups, difference in □ among groups, the number of common items, means and spreads of item discrimination and item difficulties relative to those of a total test were varied.

A comparison of different parameter estimation methods

Isabel Cañadas, Sonia Tirado and Rebeca Bautista

At present there are a wide variety of software to perform both calibration and parameter estimation of the items based on Item Response Theory (IRT). However, they all don't lead to the same results due to various issues. First, the type of algorithm used for estimation. Moreover, the own software particularities. From a bank of multiple choice items that measure knowledge of statistics the aim of this work is to analyze the differences found in the results of the item parameter estimation, following the 3-p model, depending on two aspects: a) the type of software and b) the estimation algorithm.

Omitted responses: The effect on item parameters

Sonia Tirado, Rebeca Bautista and Isabel Cañadas

In the calibration of item banks for the creation of Computerized Adaptive Tests (CAT) the presence of omitted responses from the subjects is a common occurrence when data collection is performed by pencil and paper procedures. This may affect parameter estimation results. From a bank of multiple choice items that measure knowledge of statistics, we adjusted the three-parameter logistic model (3-p) and analyzed the effect of omitted responses on the items parameters estimated by different methods.

The sample size effect on item parameters

Rebeca Bautista, Isabel Cañadas and Sonia Tirado

The sample size in item bank calibration for the creation of Computerized Adaptive Tests (CAT) may affect the parameter estimation results. From a multiple choice item bank that measure knowledge of statistics we adjusted three-parameter logistic model and analyzed the effect of different sample sizes on the items parameters when performing estimation by different methods.

Effects of medium on writing assessment of learners of chinese as a foreign language

Yu Zhu (Ph.D.) and Andy SL Fung (Ph.D.)

Hanyu Shuiping Kaoshi (HSK), the official Chinese language proficiency test, has been applying paper and pencil as the sole medium of its writing tests. To find out if the medium would have any effects on foreign learners' performance on the writing assessment, a sample of intermediate learners of Chinese will take 2 comparable HSK writing topics, one to be done with computer and the other on paper. Keystroke logs, think-aloud protocols, and in-depth interviews will be used to collect data. Findings and discussions will provide evidence and implications that may contribute to the improvements of the HSK writing test.

Estimating item response theory based item parameters when response matrix is quite sparse

Vipin Chilana

Many high-stakes exams face the issue of missing data in IRT calibrations. Given that CAT testing programs begin with calibrations of fixed-form tests to obtain IRT parameters, it is important to appropriately deal with missing data issues. High stake exams, with negative marking for wrong answers and test taking time constraints convert the cognitive assessments to speeded - tests rather than pure performance tests. This results in large items being not responded intermittently and at the end of the test. There are many factors for this missing data and some of them are non-random ones. While a good amount of research has already been done in this area, the specific way it can be dealt with would also depend on the region, uniqueness of selection decisions, peculiarities of test taking behavior and the item types. The study will evaluate these aspects in typical **Indian context** of conducting large scale high stake paper pencil tests and will attempt to estimate IRT parameters and simulate CAT using these parameters.

Development of a physical function CAT

Man Hung, David J.Weiss and Charles L. Saltzman

In an effort to address some of the long recognized problems with patient reported outcomes measures, we intend to develop and validate a computerized adaptive testing (CAT) instrument that can efficiently and precisely measure patients' physical functioning. We conducted a prospective study of 500 adult patients who presented for care of lower extremity musculoskeletal problems. Each patient completed a questionnaire with seven demographic items and 124 Patient-Reported Outcomes Information System physical function items. Each physical function item was then calibrated using a 2-parameter IRT model. A post-hoc simulation of CAT was further performed to assess varying CAT parameters.

Constructing a new fluid intelligence test: An IRT approach

Romanelli R., Saggino A. and Weiss D. J.

The aim of this research is to study the psychometric characteristics of a new fluid intelligence test, starting from Carroll's three stratum theory (1993). This test was administered to a sample of 659 Italian student. All items (n=220) were divided in two parallel forms, that were administered to two groups. A subsample (n=460) was also administered the Raven's Advanced Progressive Matrices in a counterbalanced order.

Our findings showed that the two test forms had good internal consistencies ($KR20_1$=,89; $KR20_2$=,85). For each item were computed IRT parameters according to the 3-parameter logistic model. Also CAT simulation studies were conducted.

12.00 - 13.30 Lunch

13.30 – 15.05 Paper Presentations

Multidimensional CAT

A general multidimensional computerized adaptive testing platform in health outcomes measurement

Seung W. Choi, Richard C. Gershon, and Dave Cella

Most CATs in health outcomes measurement to date are unidimensional, measuring one construct at a time. Clinical studies often involve multiple, interrelated constructs (e.g., anxiety and depression, fatigue and physical function) as outcome variables; however, conventional unidimensional approaches generally ignore or underutilize the information present between the constructs. Reducing patient burden from overtesting is critical in health outcomes measurement and has motivated exploring multidimensional approaches as a means to achieve higher measurement efficiency and brevity as a result. This presentation will discuss and demonstrate a general CAT algorithm, suitable for various multidimensional frameworks, including multi-unidimensional, bifactor, and fully multidimensional approaches.

Benefits of multidimensional adaptive testing from a practical point of view

Ulf Kröhne & Ivailo Partchev

This talk will focus on benefits of multidimensional adaptive testing (MCAT) and multidimensional scoring (MS) for three one-dimensional fixed-length computerized adaptive tests (CAT). With the help of a simulation study we will compare the following approaches: Separate CATs of the different traits with optimized starting points, a combined CAT with content control, separate CATs with MS, a combined MCAT with mixed dimensions and a combined MCAT with items ordered by dimensions. Although MCAT achieves highest efficiency, MS of separate CATs leads to empirically relevant improvements in the measurement. Some side effects will be discussed, for instance, for item exposure.

The bifactor model and its application to multidimensional computerized adaptive testing

Dong Gi Seo and David J. Weiss

The present study extends the work on bifactor CAT of Weiss & Gibbons (2007) in comparison to a fully multidimensional bifactor method using multidimensional maximum likelihood estimation and Bayesian estimation for the bifactor model (MBICAT algorithm). Although Weiss and Gibbons applied the bifactor model to CAT (BICAT algorithm), their methods for item selection and scoring were based on unidimensional IRT methods. Therefore, this study investigated a fully multidimensional bifactor CAT algorithm
using simulated data.

Multiple-Category classification with multidimensional adaptive testing in large-scale assessments

Nicki-Nils Seitz and Andreas Frey

The study demonstrates the hypothetical use of multidimensional adaptive testing (MAT) with a confidence interval classification approach for a large-scale assessment situation. Varying the location and the number of cut points, MAT is contrasted with fixed length testing (FIT) and unidimensional computerized adaptive testing (CAT). In a real data simulation, we used the ability distribution and the item parameters obtained from a calibration study of a test measuring the attainment of the German educational standards in mathematics. Since the efficiency of classification is generally dissatisfying for multiple categories, for pass/fail situations the measurement efficiency can be increased especially by MAT.

CAT for Classification

Adaptive sequential mastery testing using the Rasch model and Bayesian sequential decision theory

Hans J. Vos and Cees A.W. Glas

In this paper, a version of adaptive sequential mastery testing (i.e., classifying students as a master/non-master or to continue testing and administering another item or testlet) is studied where response behaviour is modelled by an item response theory (IRT) model. Firstly, a general theoretical model will be sketched that is based on a combination of Bayesian sequential decision theory and item response theory. Then it will be pointed out how IRT-based sequential mastery testing can be generalized to adaptive item and testlet selection rules, that is, to a situation where the choice of the next item or testlet to be administered is optimized using the information from previous responses.

The performance of IRT-based adaptive and adaptive sequential mastery testing (ASMT) will be studied in a number of simulations as a function of the proportion correct decisions and the average loss using the Rasch model. It was found that adaptive sequential mastery testing does indeed lead to a considerable decrease of loss, mainly due to a significant decrease of testlets administered. The number of correct decisions remains relatively stable. The decrease of loss is positively related to the number of items in a testlet: the larger the number of testlets and the smaller the number of items in a testlet, the less the loss. The reduction of loss due to adaptive testlet selection is less pronounced. Finally, the possibilities and difficulties of application of the approach in the framework of the 2-PL and the 3-PL will be discussed.

Effects of different termination criteria on classification consistency in CAT

Nam Keol Kim

This study is trying to investigate the effects of different termination criteria on the performance of CAT administration in terms efficiency that can be measured by not only usual evaluative index such as root mean square difference (RMSD) but also classification consistency index (CCI) which is a more practical approach, for instance, with standardized accountability achievement test in US. Through the post-hoc simulation utilizing POSTSIM3 program, the efficiency of different termination criteria has been investigated with different conditions along with different estimation methods and different IRT models applied. Applied were real response data that consisted of 2889 students on 80 mathematics test items and the item parameter were estimated through BILOGMG in R.

Using the sequential probability ratio test when items and respondents are mismatched

Maaike van Groen & Angela Verschoor

Computerized classification tests (CCT) select items adaptively for each examinee in order to make a classification decision. A decision is made regarding the assignment of an examinee into one of the mutually exclusive categories along the ability scale. The sequential probability ratio test (SPRT) can be used for making the decision.

In CCT, items are selected that provide maximum information at the current ability estimate or at the cutting point. This implies that items have to be available whose difficulty matches the examinee ability. The effect of a mismatch between difficulty and ability in tests using the SPRT is unknown.

Nominal error rates in computerized classification testing

Nathan A. Thompson

A common finding in computerized classification testing (CCT) research, though often not explicitly noted, is that observed error rates do not always follow nominal error rates, and in some cases are substantially different. There are likely several contributing factors, including the size of the indifference region, which is typically selected arbitrarily rather than empirically, and the information structure of the item bank. This paper will utilize monte carlo simulations to manipulate these two variables, and investigate their effect on the observed error rates in CCT. Both approaches to the likelihood ratio criterion (point and composite) will be utilized.

CAT in Education

Orthography testing and CAT - on the way to establish individual orthography learning

Sarah Frahm

In order to facilitate individual orthography learning, diagnostic tests are indispensable. They do not only need to give a differentiated insight into the orthography competence and its development, they also need to allow further learning with individualized learning material which is adapted to the students' individual needs.

At the IACAT, an overview of a research project of the National Education Panel Study (NEPS) in Germany with the long-term objective to develop a computerized adaptive test will be presented. It includes the development of a computerized analysis software and a mode-effect-study.

Adapt your listening test to your level

Klaas Schreuder, Angela Verschoor and Niels Veldhuijzen.

Cito produces every year tests of listening comprehension for students of 15+. An item pool comprising 315 items from tests for Dutch as a mother tongue has been calibrated using the two parameter logistic model. Items are clustered in testlets of 2 to 14 items belonging to one source program. The item pool consists of 43 testlets. In an adaptive test, the testlets should be selected as a whole. Simulations show that the length of the tests of 55 items can be reduced to approximately 40 items without increase in measurement errors.

WISCAT-pabo: Computerized adaptive testing of arithmetic skills of first-year students at primary school teacher training colleges.

Gerard J.J.M. Straetmans & Theo J.H.M. Eggen

WISCAT-pabo is the name of a computer-based adaptive testing package that is used to assess the arithmetic skills of first-year teacher training college (TTC) students in The Netherlands. There are several reasons why it was decided to develop an adaptive test. The most important among them was the fact that the population of first-year TTC students is highly heterogeneous in terms of arithmetic skills, due to the differences in previous education.
In addition to the argumentation for adaptive testing, we will address the following topics successively:

- Design of the testing package;
- Measurement quality;
- Operational results;
- Student experiences.

DESIGNING & Using adaptive tests for large scale formative assessment

John Winkley

This presentation is concerned with the development and trialling of sets of adaptive computer-based tools used primarily to assess the literacy, language and numeracy skills of adult learners in both academic and work settings in the UK.

These tools have been used extensively in thousands of centres for diagnostic and formative assessment purposes and are one of several outputs from an ongoing development of computer-based adaptive formative assessment tools.

The paper looks at the benefits and problems associated with developing, trialling and deploying these types of assessments. We consider trialling and reliability measurement, as well as learner and teacher views.

15.20 – 16.30 Paper Presentations

Item selection (Part 2)

Item selection criteria in CAT based on predictive distributions

Mathilde Huisman-van Dijk and Frans Kamphuis

In contrast to the popular item selection criteria based on Fisher information or Kullback-Leibner information the main purpose of this article is to present some new criteria for item selection in CAT, founded in a Bayesian decision theoretical perspective. The first criterion is based on predicted outcomes. The moments of the posterior predictive distribution, offer us excellent opportunities to state some selection rules. A second selection rule uses the information from the future outcomes, the expected Shannon information will be maximised. The sequence of selected items is also different in this approach in contrast to the more traditional approaches.

Content control with the maximum priority index method in multidimensional adaptive testing

Andreas Frey, Ying Cheng and Nicki-Nils Seitz

Two generalizations of the maximum priority index method (MPI; Cheng & Chang, 2009) to multidimensional adaptive testing (MAT) are introduced. The performance of the two methods is compared to unrestricted MAT. As independent variables, several characteristics of the item pool and the requested ratios of items per dimension are manipulated. MAT is carried out using the bayesian item selection and ability estimation proposed by Segall (1996) assuming knowledge of the a-priori variance-covariance-matrix. The study provides (a) easy accessible solutions for how to use content control in MAT and (b) results enabling practitioners to choose a feasible content control method.

The maximum priority index for generalized partial credit model in computerized adaptive testing

Zeng Lingyan

Item selection strategy(ISS) is the most important part of computerized adaptive testing (CAT), whose quality is directly related to the reliability,efficiency,and security of the test. Many researches and applications of CAT are based on a dichotomously scored model. It is highly evident that more information can be obtained from examinees using a polytomously scored model rather than a dichotomous model.Moreover,it is necessary for us to further explore CAT research based on a polytomously scored model. Considerable research is already being conducted on CAT using the Grade Response Model (GRM);however,in our country,there are few reports pertaining to research on CAT using the Generalized Partial Credit Model(GPCM).In the GPCM,each item contains several step parameters,and there are no specific rules governing them.The posterior step cannot advance when the earlier step has not been completed,and the posterior's step parameter may be lower than that of the previous one.The traditional maximal item information (MII) approaches could easily result in great exposure of highly distinct items.This study investigated a new heuristic approach,the maximum priority index(MPI) method using the GPCM through Monte Carlo simulation. Research shows that MPI method is better than the MII approach in such index: recovery, stability, the average item number and exprosure.

Conformation testing and assessment of change

Unproctored internet test verification: Using adaptive confirmation testing

Guido Makransky and Cees Glas

Unproctored Internet testing (UIT) is commonly used in employment test administration. This article proposes and compares a fixed length and an adaptive method for detecting if a test-taker's original UIT responses are consistent with the responses from a follow-up confirmation test. Simulation studies indicated that the adaptive confirmation test was almost four times shorter while maintaining the same detection power. The study also demonstrated that cheating can have a detrimental effect on the validity of a selection procedure, and illustrated that the use of a confirmation test can remedy the negative effect of cheating on validity.

Accepting the null: Determining no change within adaptive measurement of change

Steven W. Nydick and David J. Weiss

This paper proposes and evaluates a method of determining "no change" within change CAT. Instead of no change quantified as a "point" on a distribution with CAT testing significant deviations from that point, it might be more intuitive to think of no change as a range of values. In this context, the CAT implementer (or researcher) could decide in each implementation how they want to quantify no change depending on measurement purposes. As a first step, this paper assesses a moving classification CAT, with classification determined by a Sequential Likelihood Ratio Test, and where the classification categories at the second time point are determined from the estimated theta at the first time point.

Measuring individual growth curves with computerized adaptive testing

Shannon M. Von Minden and David J. Weiss

The measurement of individual change across multiple time points is a controversial topic riddled with psychometric problems. However, there are situations in which the measurement of individual growth curves is important. A recently completed simulation study demonstrated that conventionally constructed and scored tests cannot measure individual growth at multiple time points; IRT scoring improved the recovery of true growth curves, but there was still a wide range of error in recovery. The current study implemented CAT as a potential solution for the measurement of individual growth curves, with the focus on how well the CATs captured true change in comparison to both conventionally scored and IRT scored conventional tests.

Collatoral information in CAT

A Bayesian approach for introducing empirical information in CAT

Mariagiulia Matteucci

In this work, a Bayesian approach based on Markov chain Monte Carlo (MCMC) for the introduction of empirical prior information in the ability estimation within computer adaptive testing (CAT) is investigated. When covariates related to respondents are available and a significant relation with ability is identified, it is possible to include collateral information both in the initialization and in the ability estimation. As a consequence of the empirical initialization, item over-exposure is reduced. Furthermore, abilities are estimated with increased precision, especially for short tests. By using both simulated and real data, it is shown the advantages of introducing empirical information about the candidates for increasing the measurement precision and reducing the test length.

Text classification in prior selection of CAT

Qiwei He

To ensure the precision and efficiency of computerized adaptive testing (CAT), a prior selection is sometimes added before the implementation of the tailored-test. The typical methods of prior selection include filtering participants by demographic information and scaling scores from the previous tests. However, the information from textual components, such as writing or verbal reasoning, has been paid little attention to in this process. The major reason is probably the complex scoring model and the flexible language patterns. The central message of this paper is to demonstrate that the automated scoring algorithms in text classification can be not only useful in information extraction but also a good alternative in the prior selection for CAT.

The context of the project is the development of an alternative intake procedure for patients with post traumatic stress disorder (PTSD) with the aim to reduce the diagnose time and to relieve the workload of the psychiatrists involved. Instead of conducting face-to-face interviews with traditional questionnaires, respondents were asked to write down their stories online. Based on a collection of 200 respondents' self-reports, textual classification models were constructed by text mining techniques via two approaches, Naive Bayes Classifier and Decision Trees, within the field of machine learning. The outcome of a text classification model was used as the filter in the prior selection for the further IRT-based computerized adaptive tests.

Robust item selection in computerized adaptive testing

Bernard Veldkamp

Large scale applications of computerized adaptive testing, high costs of item calibration, in combination with exposure control problems have created favorable circumstances for the implementation of automated item generation in operational CATs. Regression methods and cloning algorithms have been proposed to predict the psychometric parameters of the generated items based on collateral information about the construction of the items, the cognitive models underlying automated item generation, the word usage, etc. But even though good results have been obtained, the values of the parameters can only be predicted with considerable standard error of measurement.

Unfortunately, most item selection rules preferably select items with high discriminating power. In other words, they focus on items with overestimated or over predicted discrimination parameters. This problem even occurs in most operational item banks, where the discrimination parameters are estimated with error. In this paper, robust item selection rules will be proposed that have been developed for dealing with uncertainties in the item parameter estimated. Different approaches will be compared and the item selection rules will be applied to an operational computer adaptive intelligence test.

17:00 **Social activity Sponsored by Cito**
(Pitch & Putt at number 22 on map)

19:00 **Dinner**

Wednesday, June 9th

09.00 - 10.15 Keynote speakers

Bootstrapping an item bank

Cees. A. W. Glas and Guido Makransky

An accurately calibrated item bank is essential for a valid computerized adaptive test. However, in some settings, such as in occupational testing, there is limited access to test takers for calibration. As a result of the limited access to possible test takers, collecting data to accurately calibrate an item bank is usually difficult. In such a setting, the item bank can be calibrated online while in operation. We explore 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 actual test takers are processed and the scores they obtain have consequences. A simulation study is used to identify the optimal calibration strategy. Manipulated variables were: the calibration strategy, the size of the calibration sample, the size of the item bank, and the item response model.

Computerized classification testing to predict an observable outcome

Matthew Finkelman, Yulei He, Giles Hooker, Wonsuk Kim, Robert Keller and Barbara Gandek

Computerized classification testing (CCT) is a well-known psychometric approach to categorizing examinees into groups. Most existing CCT methods rely on the latent variable models of item response theory (IRT), which were originally developed for assessments where there is no external outcome variable. This research examines the use of CCT in a different context: the prediction of a medical event from responses to a health questionnaire. Because the outcome is observable rather than latent, new item selection procedures and variable-length stopping rules are needed. Methods are illustrated using data from the Medicare Health Outcomes Survey.

10.15 - 10.45 Break

10.45 - 12.20 Paper Presentations

Item bank development

Pushing the limits of an adaptive test: Interaction between the test population and the item pool

Steven L. Wise, Brian Bontempo, and G. Gage Kingsbury

This study describes practical and technical issues in expanding the range of measurement for an adaptive test while the test is being used operationally. This study details the theoretical approach and the experience of adding content that is appropriate for measuring students in kindergarten through second grade to reading and mathematics test originally designed to measure students in third grade and beyond. While a great deal of research has been done concerning creating an item pool for adaptive testing, less is known about extending test content beyond that designed for the initial testing population.

Extension of content within an item pool gradually is fairly well understood. Almost every operational adaptive test has mechanisms to improve the item pool by addition of items to measure extreme students better, or to reduce item exposure. The difference here is that the amount of change in the item pool was expected to be quite large, and the change in the concepts of reading and mathematics between kindergarten and grade three are huge.

This study examines field testing procedures that were developed for the project, and evaluates their utility. It also examines processes used to examine dimensionality of the construct using information from the adaptive test. After several false starts, the project was successful in extending the measurement scale and the content of the item pool to measurement of very young students.

Optimal calibration designs for computerized adaptive testsing

Angela Verschoor

Standardized tests heavily depend on gathering data through pretests. This observation holds even more for a computerized adaptive test (CAT). Large item banks are built and calibrated, while measurement errors in a CAT depends on the precision with which the items are calibrated. In those circumstances, more items have to be sampled than one single candidate can take. Data is usually gathered through a design of linear test booklets. Typically, these designs are made by hand according to some design principles.

In this paper we investigate the efficiency of various designs and proposes an optimization method for calibration designs. The solutions of this method are compared to designs constructed according to the 'traditional' best-practice principles. Simulations show that optimal designs have approximately 5% lower standard errors for the item parameter estimates. Based on this experience, practitioners might decide to improve their parameter estimates or to reduce the sample size by 10% in order to achieve estimation errors in the same order of magnitude as before. The latter option would typically imply cost reduction of 10% in the logistics of the pretest campaign for the development of a new CAT.

In a large-scale School Achievement Test taken by 70% of the grade-7 pupils (age 11) in the Netherlands, 120 items are taken on Arithmetic. Responses of 100,000 pupils to these Arithmetic items were used to make comparisons between various designs. A similar improvement in estimation accuracy has been found for the optimal designs.

Flexibly and accurately evaluating item pool adequacy and optimization for linear-to-CAT conversion using optimization

Xiang Bo Wang and Xinhui Xiong

Using optimization programming language (OPL) of CPLEX 11.0 (ILOG, 2007), this study is to demonstrate the power and flexibility of using linear models (van der Linden, 2005) in assessing the adequacy and optimization of the item pool of a well known international P&P assessment for a planned CAT conversion, while providing the optimal solutions satisfying all content, psychometric, and item exposure requirements. Both the weaknesses and strengths of the item pool will be pointed out, and methods of alleviation for the pool weaknesses and of capitalization of the pool strengths through optimization will also be detailed.

Applications of CAT (part 2)

A comparison of item exposure, test information, and theta recovery for LOFT, CAT, and fixed form tests.

Kirk A. Becker, Brian Bontempo, Phil Dickison, and James S. Masters

This paper presents research comparing the characteristics of tests administered under different selection algorithms. Several operational testing programs are used to compare linear on-the-fly (LOFT), computer adaptive (CAT), and multiple fixed forms. Item pools are calibrated with the Rasch (1PL) IRT model, and include both dichotomous and polytomously scored item pools. Item usage, theta recovery, decision consistency, test information and standard error, and average item overlap between simulees are compared for the three administration options. This research will provide practical advice on the selection of different test designs relative to program and item pool characteristics.

Oppertunities on adaptive testing in India

Dr. V. Natarajan and Mr. Rajeev Mennon

In this presentation, in the slot given to us by the conference, the authors Madan Padakki, CEO and Dr. Venkatesa Natarajan, Prof, Emeritus, MeritTrac Services Pvt. Ltd. would like to bring to the group the immense opportunities that lay before assessment companies conducting recruitment test to several national and international clients and also several institutions of higher learning for their admission tests and semester examinations. There are several testing organizations, some of them exclusively conducting entry level recruitments for various companies, though some of them combine training, placement and recruitment examinations. MeritTrac conducts exclusively recruitment examinations and assessments for educational institutions. This presentation consists of two parts, the first part presenting activities, innovations and R&D activities involving both offline (paper-pencil test) and online examinations, of late, particularly the past and present activities of MeritTrac services Pvt. Ltd. The concluding part brings the immense opportunities that lay before the country in terms of computerized online and computer adaptive testing including experiments with adaptive testing both online and offline modes. Also some glimpses of R&D achievements of producing Ph.D level work related to Item Response Theory (IRT) and Adaptive Testing.

A dynamic way of equating for cat practice

Zhang Quan

The paper presents a dynamic way of equating better suited to today's practice of computerized adaptive testing. Beginning with a brief description of equating via BILOG for paper-and-pencil tests, the author would demonstrate a way of equating (with simulated data), in which the linking items are randomly demonstrated while test takers cope with each item. The author believes that, with rapid development of computer technology, such a practice has been feasible and turns out to be more efficient than the item-bank supported CAT practice. The author has been getting involved in language testing on large scale in China ever since 1989 and would like to have such an idea as a basis for further discussion with language testing counterparts abroad.

Graduate Management Admission Council® | Assessment Systems Corporation www.assess.com | NWEA Northwest Evaluation Association Partnering to help all kids learn | MeritTrac™ India's Largest Skills Assessment Company | CollegeBoard inspiring minds™ | cito | pi Company | RCEC Research Center voor Examinering en Certificering | IBR Institute for Behavioral Research

Item Exposure

Conditional multiple maximum exposure rates in CATs

Juan Ramon Barrada, Julio Olea and Francisco J. Abad

Item bank security is a major concern in computerized adaptive testing. A common method for improving security is to impose a maximum exposure rate (rmax) that no item should surpass. For doing so, as many item exposure parameters as items composing the bank (n) are calculated. Recently, a new approach, the multiple-rmax method, was proposed: instead of using a single value of rmax, as many rmax values as items to be presented (Q) are used. In this way, n*Q exposure control parameters are used. By doing so, important improvements in item exposure control can be achieved, with minor decrements in accuracy. A problem with this proposal is that is not guaranteed that the item exposure rates will be below the limit when considering examinees of the same or proximal ability. The idea of controlling exposure rates conditional on ability rates has been already explored with the methods of single rmax value. In this case, the continuum of the ability distribution is discretized is K points (usually, 10-12) and different sets of exposure control parameters are calculated for each of these K points. In this study, we will show how to expand the multiple-rmax method for the case that control is desired conditional on ability, how to compute the n*Q*K exposure control parameters and the main results of several simulations where the performance of this method was evaluated.

The on-line procedure for simultaneous control of item exposure and test overlap in variable length computerized adaptive testing

Chia-Ling HSU and Wen-Chung WANG

Hsu and Chen (2007) proposed a modified SHT procedure (SHTV) to simultaneously control item exposure and test overlap in variable length computerized adaptive testing. Unfortunately, SHTV requires time-consuming simulations and it is dependent of simulation setting. To resolve these problems, we developed an on-line procedure of SHTV, called SHTVO, and evaluated its performances. The results show that (a) both SHTVO and SHTV could maintain control of item exposure and test overlap, but SHTVO outperformed SHTV when conditions were stringent; (b) when real settings and simulation settings were different, SHTV failed to control item exposure and test overlap.

Controlling on-line item exposure and test overlap in CAT with the ability-based guessing model

Shiu-Lien Wu, Wen-Chung Wang and Shu-Ying Chen

It is reasonable to assume that the guessing process concerns with person's ability rather than item property when subjects don't know the correct answer for multiple-choice tests. The purpose of our study is to inquire whether the Sympson and Hetter procedure with test overlap control (SHTOR) still performs well by using the ability-based guessing (AG) model instead of 3PL in CATs. Results show that the SHTOR performs best among four on-line test security control methods by using the AG model, as the results of Chen, Lei, and Liao (2008) by using 3PL. Thus, AG model is feasible choice in CATs.

12.00 - 12.15 **Closing of the IACAT conference**

12.15 - 13.00 **Lunch**

UNIVERSITY OF TWENTE.

IBR RESEARCH CENTRE FOR EXAMINATION AND CERTIFICATION CITO now you know

Announcement

1st RCEC Workshop on IRT and Educational Measurement

The 1st Workshop on Item Response Theory and Educational Measurement will be held at the University of Twente from Tuesday, October 12 through Thursday October 14, 2010.

The keynote speaker at this workshop will be Norman Verhelst of Cito in Arnhem (NL).

More information you can find on www.rcec.nl/irtworkshop.

Graduate Management Admission Council
Assessment Systems Corporation www.assess.com
NWEA Northwest Evaluation Association Partnering to help all kids learn
MeritTrac India's Largest Skills Assessment Company
CollegeBoard inspiring minds
CITO
pi Company
RCEC Research Center voor Examinering en Certificering
IBR