International Association for Computerized Adaptive Testing

IACAT

Advancing the Science and Practice of Human Assessment

IACAT 2015 Abstracts Booklet

Queens' College, University of Cambridge, England – September 14-16, 2015

Hosting Sponsor

The Psychometrics Centre, University of Cambridge (Prof John Rust: Director)

Organizing Committee

Prof. John Rust, Director, Psychometrics Centre

Dr. David Stillwell, Deputy Director Psychometrics Centre

Yin Wah Fiona Chan, Psychometrics Centre

Dr. Michal Kosinski, Stanford University

Suzy Howes, Suzy Howes Associates

Charlie Howes, Suzy Howes Associates

Scientific Committee

John Barnard, Professor Exec Director: EPEC Pty Ltd President, IACAT

Cliff Donath, Global Certification Program Manager: Philips Healthcare Exec Director, IACAT

Theo Eggen, Professor: University of Twente Psychometrician CITO Board member, IACAT

Kathi Gialluca, Senior Research Scientist,Perason Treasurer, IACAT

Tetsuo Kimura, Professor, Niigata Seiryo University, Board Member, IACAT

Gage Kingsbury, Consultant, Psychometric Consulting

Lawrence Rudner, Consultant Past President, IACAT

Nate Thompson, Vice President ASC Membership Director, IACAT

David Weiss, Professor: University of Minnesota President Emeritus, IACAT

Duanli Yan, ETS Secretary, IACAT

TABLE OF CONTENTS

CONFERENCE PROGRAM

Monday 14 September 2015

Registration 08:00 – 16:30

TIME	BOWETT ROOM	ARMITAGE ROOM	ERASMUS ROOM	ANGEVIN ROOM
08:45 – 10:15	Workshop 1 **Computerized Multistage Adaptive Testing** Duanli Yan, Alina von Davier and Kyung "Chris" Han (*GMAC*)	Workshop 2 **CAT simulations: how and why to perform these?** Presenter: Angela Verschoor and Theo Eggen (*Cito and University of Twente*)	Workshop 3 **An introduction to CAT** Presenter: Nate Thompson (ASC)	Workshop 4 **Building and delivering online CAT using open-source Concerto Platform** Presenter: Michal Kosinski (*Stanford*)
10:15 – 10:30	Break			
10:30 – 12:00	Workshops continued			
12:00 – 12:45	Lunch			
12:50 – 13:00	**Opening and Welcome** John Rust (*Psychometrics Centre*)			
13:00 – 13:30	Presidential address **Title: Improving precision of CAT measures** John Barnard (*EPEC*)			

13:30 – 14:10	Keynote presentation 1 **Topic: A self-replenishing Adaptive test** Wim van der Linden (*Pacific Metrics Corporation*) Chair: John Barnard			
14:10 – 14:30	**Break**			
	Parallel Session 1 (Symposium) Chair: Alina von Davier	Parallel Session 7 (Symposium) Chair: Shiyu Wang	Parallel Session 13 (Symposium) Chair: Gage Kingsbury	
14:30 – 16:00	**Topic: CAT around the world** Presenters: Tetsuo Kimura: CAT developments in Japan John Barnard: CAT Down Under – developments in Australia Marie de Beer: CAT in South Africa John Rust: CAT in the United Kingdom	**Topic: New developments in CAT by graduate students at UIUC** Presenters: **Study 1** (Hyeon-Ah Kang, Yi Zheng & Hua-Hua Chang): Online Calibration Strategies for a Joint Model of Item Responses and Response Times in CAT **Study 2** (Justin Kern & Hua-Hua Chang): Maximizing All the Information: Using Response Times in CAT **Study 3** (Edison M. Choe, Jinming Zhang & Hua-Hua Chang): Utilizing Response Time in Sequential Detection of Compromised Items in CAT	**Topic: Some bothersome problems for operational CAT and some potential solutions** Presenters: Anthony Zara: Producing CAT scores for test takers who run out of time Brian D. Bontempo: CAT for Rapidly Changing Content Domains G. Gage Kingsbury: Can CAT give us knowledge about idiosyncratic performance?	
16:00 – 16:15	**Break**			

	Parallel Session 2 (Symposium continued) Chair: Alina von Davier	Parallel Session 8 (Symposium continued) Chair: Shiyu Wang	Parallel Session 14 Chair: Shungwon Ro	
16:15 – 17:45	**Topic: CAT around the world - continued** Presenters: Andreas Frey: CAT in German speaking countries Theo Eggen: CAT in Education in The Netherlands Mariana Curi: CAT developments in Brazil. Alina von Davier: Adaptive tests in the USA	**Topic: New developments in CAT by graduate students at UIUC - continued** **Study 4** (Chanjin Zheng & Hua-Hua Chang): Stratification strategies for CD-CAT based on linear/binary search **Study 5** (Susu Zhang & Hua-Hua Chang): The Relationship between Q-Matrix Specification and Item Exposure Rate in CD-C	**Topic: A computer-adaptive measure of delay discounting** Presenter: Vaishali Mahalingam, Michael Palkovics, Michal Kosinski & David Stillwell **Topic: Improving CAT test fairness with comparable latency for pretest items** Presenter: Sung-Hyuck Lee, Bruce Williams, Wugen Dai, Richard Sullivan & Jason He **Topic: Computer adaptive measurement of quality of life across cultures: Results from IRT and CAT simulation studies** Presenter: Chris Gibbons **Topic: A Variable Length CAT for Ranking Data in Forced Choice Assessments** Presenter: Shungwon Ro, Chia-Wen Chen, Wen-Chung Wang and Xue-Lan Qiu	
18:30	Welcoming reception: Fitzwilliam museum			

Tuesday 15 September 2015

09:00 – 09:40	Keynote presentation 2 **Title: Multidimensional CAT: Calibration, model fit and secondary analysis** Cees Glas (*University of Twente*) Chair: Theo Eggen		
09:40 – 10:15	Early Career Researcher Award **Title: Optimal Design and Scoring for Adaptive Multi-Stage Testing: A Tree-Based Approach** Duanli Yan (*ETS*) Chair: John Barnard		
10:15 – 10:30	**Break**		
	Parallel Session 3 Chair: Nate Thompson	Parallel Session 9 Chair: Haniza Yon	Parallel Session 15 Chair: Marie de Beer
10:30 – 12:00	**Topic: A practical model for CAT development** Presenter: Nathan Thompson **Topic: An exploratory study of starting a CAT with a non-scaled item pool** Presenter: Deborah Harris, Chunyan Liu & Troy Chen **Topic: Moving the New Zealand Progressive Achievement Tests into the Computer**	**Topic: Development of a computerized adaptive face perception test** Presenter: Roeland Verhallen & Luning Sun **Topic: Development of an adaptive integrity test** Presenter: Haniza Yon, Norsyahida Abd Kadir, Nur Ayu Johar & Nur Mutalib **Topic: Response Formats and Trait Estimation Efficiency in Computerized**	**Topic: Evaluating the comparability of CAT tests across test delivery platforms** Presenter: Agnieszka Walczak & Ardeshir Geranpayeh **Topic: Empirical comparison of scoring rules at early stages of CAT** Presenter: David Magis **Topic: An Improved Online Item Calibration**

	Adaptive Domain Presenter: Hilary Ferral	**Adaptive Testing** Presenter: Yin Lin & Anna Brown	**Method for Multidimensional Computerized Adaptive Testing** Presenter: Ping Chen, Chun Wang & Jihong Xu
12:00 – 12:55	Lunch / IACAT Board meeting		
13:00 – 13:40	Keynote presentation 3 **Title: Some exciting new developments concerning CAT foundations and implementations** Hua-Hua Chang (*University of Illinois*) Chair: Nate Thompson		
	Parallel Session 4 Chair: Andreas Frey	Parallel Session 10 Chair: Alper Şahin	Parallel Session 16 Chair: Tetsuo Kimura
13:45 – 15:15	**Topic: Which way did he go?: An examination of routing and scoring in a ca-MST** Presenter: Andrew Dallas & Richard Luecht **Topic: Development of Multi Stage Tests based on teacher ratings.** Presenter: Stephanie Berger **Topic: Multistage Testing with routing based on performance on different test sections** Presenters: Kyung (Chris) Han, Fanmin Guo & Eileen Talento-Miller	**Topic: CloudCAT: Implementing CAT as a Web-service** Presenter: Haniza Yon, Rense Lange, Nur Ainshah, Abid Altaf & Norsyahida Kadir **Topic: The design and development of a web-based adaptive Raven's-like automatic test generator** Presenter: Isaac Thimbleby **Topic: Effects of errors in item parameter estimates on recovery of theta estimates in CAT** Presenter: Alper Şahin & David J. Weiss	**Topic: Improvement and evaluation of a small-scale ESP CAT** Presenter: Tetsuo Kimura & Yukie Koyama **Topic: An application of computerized adaptive testing in audiological assessment** Presenter: Wayne Garrison & Joseph Bochner **Topic: Estimating reliability in Linear on the fly (LOFT) designs** Presenter: Tammy Trierweiler

15:15 – 15:30	**Break**	
15:30 – 16:15	Keynote presentation 4 **Topic: The future of CAT should be open source** Michal Kosinski (*University of Stanford*) Chair: John Rust	
16:30	Excursion: College Backs Punting Tour	
18:30	Conference dinner: St Catherine’s College	

Wednesday 16 September 2015

09:00 – 09:40	Keynote presentation 5 **Topic: Happy CAT: Options to allow test takers to review and change responses in CAT** Kyung (Chris) Han (*GMAC*) Chair: Alina von Davier		
09:40 – 10:15	Keynote presentation 6 **Topic: Test construction based on the Rasch Poisson counts models** Heinz Holling (*University of Muenster*) Chair: John Rust		
10:15 – 10:30	**Break**		
	Parallel Session 5 Chair: Gage Kingsbury	Parallel Session 11 (Early career researchers Part 1) Chair: John Barnard	Parallel Session 17 Chair: Duanli Yan
10:30 – 12:00	**Topic: Comparing CATS and the block review method in providing review options in CAT** Presenter: Zhongmin Cui, Chunyan Liu, Yong He & Hanwei Chen **Topic: Identifying item enemies in a CAT pool using a chi-square goodness-of-fit test** Presenters: Brian Bontempo, Steve	**Topic: A Partial Likelihood Method for Computerized Adaptive Testing to Allow for Response Revision** Presenter: Shiyu Wang **Topic: Evaluating effectiveness of standard error of score estimation as a termination criterion in CAT.**	**Topic: Skinning a CAT in the real world: How an educational authority transitioned a low stakes assessment to CAT using Concerto** Presenter: Andrew Kyngdon **Topic: Skinning a CAT … continued (double session)** Presenter: Andrew Kyngdon

	Wise, Gage Kingsbury & Ron Houser **Topic: Optimal greedy item selection for constrained tests** Presenter: Daniel Bengs	Presenter: Chansoon Lee **Topic: Rules induction based method for the item selection in computer adaptive testing.** Presenter: Maria Rafalak **Topic: Latent Class Based Item Selection for CAT in Progress Tests.** Presenter: Nikky van Buuren	**Topic: Heuristic Constraint Management Methods in Multidimensional Adaptive Testing** Presenter: Sebastian Born	
12:00 – 12:55	Lunch			
	Parallel Session 6 Chair: Richard Luecht	Parallel Session 12 (Early career researchers Part 2) Chair: John Barnard	Parallel Session 18 Chair: Alper Şahin	
13:00 - 14:30	**Topic: Transitioning from linear testing to Multistage Testing: A case study.** Presenter: Maaike van Groen **Topic: Design and implementation of a large-scale computer-3-4 zed adaptive multistage testing system for reading and listening** Presenter: Richard Luecht **Topic: The development of a computerized adaptive test with online**	**Topic: Self-Adapted Testing as Formative Assessment:** **Effects of Feedback and Scoring on Engagement and Performance**. Presenter: Meirav Arieli-Attali **Topic: Weighted-Probability Based Classification for Computerized Classification Testing.**	**Topic: Multidimensional IRT versus Mean Adjustment in Estimating Cognitive Ability Scores** Presenter: Darrin Grelle **Topic: The first stage items selection methods of CD-MST**	

	calibration Presenter: Angela Verschoor & Stephanie Berger	Presenter: Victoria Song **Topic: Item selection criteria for Logistic Positive Exponent model-based Computerized Adaptive Testing** Presenter: Thales Akira Matsumoto Ricarte **Topic: Application of the Adaptive Measurement of Change to Measuring Achievement Growth in Reading and Mathematics** Presenters: Chaitali Phadke, David J. Weiss & Theodore Christ	Presenter: Chunlei Gao, Zhaosheng Luo, Chanjin Zheng, Xiaofeng Yu, Peida Zhan **Topic: Two ways to unleash CAT's true potential** Presenter: Thomas Garrard	
14:35 – 15:05	Keynote presentation 7 **Topic: Learning parameters in learning environments: Trials and tribulations** Kevin Wilson (*Knewton*) Chair: Cliff Donath			
15:05 – 15:30	**Wrap-up and closing**			

ABSTRACTS FOR WORKSHOPS

Workshop 1: Computerized Multistage Adaptive Testing

Duanli Yan, ETS, Alina von Davier, ETS, Chris Han, GMAC

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

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

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

Workshop 2: CAT simulations: How and Why to Perform These?

Angela Verschoor, CITO, Theo Eggen, CITO

In this workshop, the goals and usefulness of simulations for constructing CATs will be discussed. The measurement characteristics of a CAT can be studied and set before publishing it. Information can be collected by simulation studies that use the available IRT calibrated item bank and the proposed target population. The performance of proposed selection algorithms and constraints can be studied. Customized software will be demonstrated and distributed. Participants will practice using the software for some examples. Participants are invited to bring their own laptops for practicing (Windows®).

Workshop 3: Introduction to Computerized Adaptive Testing

Nathan Thompson, ASC

This workshop provides an overview of the primary components and algorithms involved in CAT, including development of an item bank, calibrating with item response theory, starting rule, item selection rule, scoring method, and termination criterion. It will also provide a five-step process for evaluating the feasibility of CAT and developing a real CAT assessment, with a focus on validity documentation. The workshop is intended for researchers that are familiar with classical psychometrics and educational/psychological assessment but are new to CAT.

Workshop 4: Building and Delivering Online CAT Using Open-Source Concerto Platform

Michal Kosinski, Stanford University

During this hands-on workshop participants will learn how to build and deliver an online Computerized Adaptive Test using Concerto v4, an open-source R-based adaptive testing platform. We will start with an introduction to Concerto, build HTML-based item templates, import item content and parameters and combine it all into a fully-functional online test.

ABSTRACTS FOR KEYNOTE PRESENTATIONS

Presidential Address: Improving precision of CAT measures

John Barnard, Executive Director, EPEC Pty Ltd & Professor, Universities of Sydney and Cape Town

The basic idea of adaptive testing is quite simple and has been implemented for over a century (Binet-Simon; oral examinations; etc.). Over the years item selection algorithms such as MI, MPP and WI have been developed to maximize efficiency and convergence and MLE, EAP and MAP are commonly used to estimate ability.

Dichotomously scored MCQs are mostly used to obtain response vectors in CATs. This means that a response is scored as either correct or incorrect. However, a correct response doesn't necessarily mean that the test taker knew the answer. Although the SEM increasingly decreases as the provisional ability is estimated, the question is whether the process can be improved at the item response level. In other words, can more information be extracted from a response than a simple 0 or 1? In my presentation this question is addressed.

Keynote 1: A Self-replenishing Adaptive Test

Wim van der Linden, Pacific Metrics Corporation

Items in the operational pool for an adaptive test have a restricted life span. Ideally, we should be able to replace them periodically, using the response data immediately both to calibrate the new items and score the examinees. In my presentation, I will show how a fully Bayesian approach to calibration, item selection, and examinee scoring statistics can be exploited to realize the ideal.

Keynote 2: Multidimensional CAT: Calibration, Model Fit, Secondary Analyses

Cees Glass, Department of Research Methodology, Measurement and Data Analysis, University of Twente, the Netherlands

Computerized adaptive testing is becoming more and more prominent, not only in educational measurement, but also in fields as industrial and organizational psychology and in health assessment, especially in the field of assessment of quality of life and the field of physical ability. While for educational measurement unidimensional IRT models are usually satisfactory, a field such as assessment of quality of life often calls for multidimensional IRT models. In this presentation, we outline specific problems encountered when developing a multidimensional CAT (MCAT) for fatigue for use with rheumatoid arthritis patients. The itembank consists of polytomously scored items. Some of the problems discussed apply to CAT in general, while some others are specific for MCAT.

The first topic addressed is the calibration phase. Discussed are the item administration design, and the estimation and testing procedures used. The second topic concerns the operational phase. Discussed are the item administration procedure, estimation of person parameters, models fit, differences in parameter estimates between the calibration and the operational phase, and updating strategies for item parameters. The final topic is how to contact secondary analyses using the patients' parameter estimates. Critical comments are made regarding debatable common practices and suggestions are made to improve common practices in secondary analysis using data emanating from MCAT, and CAT in general.

Keynote 3: Some Exciting New Developments Concerning CAT Foundations and Implementations

Hua-Hua Chang, University of Illinois

This presentation introduces several exciting new developments concerning Computerized Adaptive Testing (CAT) foundations and implementations. The first one is the establishment of a mathematical foundation to demonstrate that an examinee should be allowed to revise the answers to previously administered items during the course of testing. Currently very few operational CAT programs permit item revision, which is allowed by the traditional paper-and-pencil tests. This has become such a main concern for both examinees and testing companies that some testing programs have decided to switch from CAT to other modes of testing. Most recently, Wang, Felloris, and Chang have demonstrated that allowing item revision will not compromise test efficiency and security.

Then, we address a number of issues emerging from large scale implementation and show how theoretical works can solve practical problems. Our new focus will be on Cognitive Diagnostic CAT (CD-CAT), which has become a powerful tool for schools to assess students' mastery of various skills. In particular, we will present our research results on the use of CD-CAT to improve STEM learning and retention. We will also show how CD-CAT can support individualized learning on a mass scale. Lastly, we will ruminate on and discuss some possible future directions of research on CAT.

Keynote 4: The future of CAT should be Open Source

Michal Kosinski, University of Stanford

Beyond high-stake testing, the applications of CAT are still frustratingly rare. Furthermore, even well-established adaptive tests often employ only the most rudimentary methods, lagging decades behind the cutting edge of CAT research. Finally, a shortage of talent and software tools inflate the costs of expanding CAT portfolios incurred by the testing industry and its clients. As a result, individuals and economies are affected by the preventable loss of time, misallocation of talent, and decreased efficiency. I argue that the community of test publishers and CAT researchers could efficiently address these problems by embracing an open-source approach. We should focus on three areas: developing open-source research tools, open-source testing platforms, and open-source item banks. I will discuss how open-source approaches can boost CAT research, expand the pool of CAT talent, increase the quality of CAT tests, and increase profits in the CAT industry.

Keynote 5: Happy CAT: Options to Allow Test Takers to Review and Change Responses in CAT

Kyung (Chris) Han, GMAC

With its well-known advantages such as improved measurement efficiency, computerized adaptive testing (CAT) is quickly becoming mainstream in the testing industry. Many test takers, however, say they are not necessarily happy with the testing experience under CAT. Most (if not all) CAT programs do not allow test takers to review and change their responses during the testing process in order to prevent individuals from attempting to game the CAT system. According to findings from our recent research study, more than 50% of test takers complained about increased test anxiety due to these CAT restrictions and more than 80% of test takers believe they would perform better on the test if they were allowed to review and change their responses. In this keynote session, Chris Han from Graduate Management Admission Council (GMAC®) will introduce several CAT testing options that would allow for response review and revision while still retaining the measurement efficiency of CAT and its robustness against attempts to game the CAT system.

Keynote 6: CAT and Optimal design for Rasch Poisson Counts Models

Heinz Holling, University of Münster

The Rasch Poisson counts model (RPCM) may be considered as the first item response theory (IRT) model. This model allows for the analysis of count data which are assumed to be distributed according to a Poisson distribution. Although many educational and psychological tests yield such data the RPCM has gained little attention compared to other IRT models designed for binary or polytomous responses. Opposite to its counterpart the logistic Rasch model the classical RCPM as published in the monograph by Rasch (1960) does not benefit from adaptive testing strategies. However, recently developed extensions of this model will be more efficient using such procedures.

In this presentation two issues will be addressed. First, locally D-optimal designs for calibrating item parameters of a RPCM using a K-way layout with binary explanatory variables are derived. To overcome the dependence of the parameters to be estimated a sequential procedure will be introduced. The proposed method is especially suited for tests consisting of automatically generated rule-based items such as the Münster Mental Speed Test. The second topic concerns adaptive testing using the item characteristic curve Poisson counts model (ICCPCM) recently introduced by Doebler, Doebler and Holling (2014). To justify the application of this model which is more general and flexible than the classical RPCM we will introduce the covariate adjusted frequency plot.

Keynote 7: Learning Parameters in Learning Environoments: Trials and Tribulations

Kevin Wilson, Knewton

Calibration of items for adaptive testing is a very well-studied problem, and frequently the solution involves carefully curating the populations exposed to particular items. However, in an adaptive learning environment, students are exposed to items based on their current goals and needs. Thus, in adaptive learning environments, response patterns of students are necessarily biased in ways that can typically be avoided in assessment contexts. In this talk, we discuss several variations on item response theory and response patterns observed in data from Knewton’s adaptive learning platform, and we connect the pitfalls they highlight to corner cases of these models that practitioners should be aware of.

ABSTRACTS FOR INDIVIDUAL PRESENTATIONS

Design and Implementation of a Large-Scale Computer-3-4 zed Adaptive Multistage Testing System for Reading and Listening

Richard M. Luecht

Abstract:

This presentation will provide an overview of a relatively complex, web-based 1-3-4 computerized adaptive multistage testing (ca-MST) assessment system designed to separately measure English reading and listening proficiency. EFSET PLUS™ was developed to provide efficient classifications aligned to the six levels included in the Council of Europe Framework of Reference (CEFR) for languages. The testing system and content was implemented for world-wide, online delivery in about two years: from initial conception to full realization and special applications. The adaptive testing system was designed using a ca-MST "panel" configuration introduced by Luecht and Nungester (1998) and expanded by Luecht (2014). Some of the unique challenges included: (a) using mixed-format innovative item types for listening and reading; (b) designing and assembling modules and multiple ca-MST panels using modules that covered the full range of CEFR proficiencies; (c) implementing an integrated system for real-time routing; (d) standard setting and development of score scales and interpretive materials; and (e) implementing strong quality control mechanisms to monitor the panels and system performance over time.

The presentation will focus on the adaptive aspects of EFSET. Practical experiences related to the ca-MST design and associated system design modifications will also be shared.

References

Luecht, R. M. (2014). Computerized adaptive multistage design considerations and operational issues (pp. 69-83). In D. Yan, A. A. von Davier & C. Lewis (Eds.) *Computerized Multistage Testing: Theory and Applications*. London, UK: CRC Press/ Taylor & Francis Group.

Luecht, R. M. & Nungester, R. J. (1998). Some practical applications of computerized adaptive sequential testing. *Journal of Educational Measurement*, 35, 229-249.

Which way did he go? : An examination of routing and scoring in a ca-MST

Andrew Dallas. Richard M. Luecht – University of North Carolina- Greensboro Xinrui Wang – Pearson VUE

Abstract:

The computer adaptive- Multi-Stage Testing (ca-MST) mode of test administration has become extremely popular in the testing industry. Testing programs such as the GRE (Davey & Lee, 2011) and AICPA (Melican, Breihaupt, Zhang; 2009) currently employ ca-MST and others are currently exploring whether ca-MST may be of benefit to their testing programs. This paper examined the overall effects of routing and scoring within a computer adaptive multi-stage framework (ca-MST).

Testing companies enjoy the efficiency benefits of ca-MST as compared to traditional linear testing, as well as its quality-control features over computer adaptive testing (CAT). Also, test takers enjoy being able to go back and change responses before being assigned to the next module.

Lord (1980) outlined a few salient characteristics that should be investigated before the implementation of multi-stage testing. Of these characteristics, decisions on routing mechanisms have received the least attention. In the current study, routing and scoring methods were fully crossed in a 2 x 2 design to examine the impact on the ca-MST. There are two ways to select modules: (a) using a maximum-information module selection criterion to minimize the error of estimate, or (b) routing based on fixed cut score points to achieve desired item (module) exposure levels based on an assumed distribution of examinees. There

are also two scoring methods that can be applied with either one of these two module-selection methods: (i) IRT scoring using either maximum likelihood or Bayes estimates of the proficiency scores, θ; or (ii) number correct (NC) score routing using a routing table.

This paper varied both item pool characteristics (e.g., the location of information), and ca-MST configuration characteristics such as the ca-MST configuration design (e.g., 1-3, 1-2-3, 1-2-3-4). The location of the information was controlled by three levels of average item difficulty. This constructs item banks in three different scenarios. In the first scenario, the examinees outperform the items (μ_b= -1.00). This type of construction would be common in medical licensure examinations. In the second scenario, the average item bank difficulty is targeted at the mean of the examinee population (μ_b= 0.00). This type of item bank construction would be most common in the educational testing arena. Lastly, the average item bank difficulty is targeted above the examinee population (μ_b= 1.00). This type of item bank construction may be employed in an admissions type scenario where there are limited slots to enter a profession/program.

The amount of information in the bank was affected largely by the average item discrimination. The true parameters were developed to have an average discrimination of 0.6 or 1.0. Lastly, this paper varied the size of the modules (10 it vs 20 it) to observe the interaction effect of adaptivity with test length.

The results from this study shows that number correct scoring can serve as a capable surrogate for IRT calibrations at each step and that even if three-parameter scoring models are used at the end that the number correct method will not misroute as compared to traditional methods.

A Computer Adaptive Measure of Delay Discounting

Vaishali Mahalingam (The psychometrics Centre, Univ. of Cambridge), Michael Palkovics (Dept. of Psychology, Univ. of Vienna), Michal Kosinski (Graduate School of Business, Stanford Univ.), David Stillwell (Judge Business School, Univ. of Cambridge)

Abstract:

Delay discounting has been linked to important behavioural, health and social outcomes, including academic achievement, social functioning, and substance use, but thoroughly measuring delay discounting with multiple time delays and amounts is time consuming. We develop and validate an efficient and psychometrically sound computer adaptive measure of delay discounting based on a large dataset of N = 4,190 participants. Study 1 discusses the development of a binary search-like algorithm to measure delay discounting, and presents the results of a simulation study comparing the newly developed algorithm to item response theory-based computer adaptive testing and a standard static measure. Study 2 presents evidence of concurrent validity with a standard measure of delay discounting, and convergent validity with addictive behaviour and the BIS-11 questionnaire measure of impulsivity. The new measure is shorter than standard measures, includes a range of time delays, can be applied to multiple reward magnitudes, and shows excellent concurrent and convergent validity.

Development of a computerised adaptive face perception test

Roeland Verhallen & Luning Sun

Abstract:

The original Mooney Face Test comprises forty images of faces, images that consist solely of pure black and pure white elements (i.e., two-tone images). However, the original version is of limited length and unsuitable for online testing or for a test-retest paradigm. The present study created a new extended version of the Mooney Face Test, which consist of 144 three-alternative forced-choice trials, combining each face with two non-face (abstract) distractors. A total of 505 participants took part in the

test and completed all of the trials. Based on the calibrated item bank, an adaptive procedure was implemented on the Concerto testing platform. Post-hoc simulation was conducted, where various aspects of adaptive testing, including the starting point, item selection algorithm, scoring procedure and termination criterion were evaluated and discussed.

Improving CAT Test Fairness with Comparable Latency for Pretest Items

Sung-Hyuck Lee, Bruce Williams, Wugen Dai, Richard Sullivan, ACT, Inc. Jason He, University of North Carolina at Chapel Hill

Abstract:

Participants are commonly recruited in field studies in order to obtain psychometric characteristics of new items. However, such participants are generally less motivated than examinees who are administered an actual test. Thus, a big challenge for a field study is to fully engage the participants with the new items.

Rather than administering only new items in a field test, pretest items may be administered along with operational items used to compute examinee scores. If the examinees do not know which are operational or pretest items, then the examinees are motivated to fully engage with all the items. Thus, embedding pretest items in an operational test provides more validity for the statistics supporting their eventual usage as operational items.

In the past, pretest items have been selected randomly for administration without regard to the amount of time required to read, think, and provide a response to each pretest item. As a result, some examinees would be administered pretest items which require more response time. If the total testing time is fixed, then examinees who spend more time on pretest items would spend less time on the operational items which are used for scoring. In order to provide test fairness, pretest items should be chosen in such a way that no examinee is forced to spend an undue amount of time on the pretest items.

Proposed solution

Computer-based tests are commonly able to record the amount of time spent on each item. In this paper, item latency is defined as a “typical” response time (e.g., median) required for examinees to answer an item. Accordingly, the pretest latency is defined as the sum of item latencies over the pretest items administered in the test. The pretest latency is interpreted as the amount of time that is typically required for examinees to respond to the pretest items administered.

In this study, we utilize item latency as a means for selecting pretest items, in order that examinees are administered pretest items that have approximately equal pretest latency. This will prevent the time spent on pretest items from unduly influencing an examinee’s score.

Simulation and the evaluation criteria

A simulation study will be performed using empirical data. The response times for the pretest items are sampled from the empirical cumulative density functions (CDFs) obtained from the actual data set, in which a separate empirical CDF for a pretest item is computed at each of a specified number of theta intervals. In the simulation, a simulee’s response time for a particular administered pretest item is randomly sampled from the empirical CDF for that item conditioned on the simulee’s true theta.

The initial item latencies for the pretest items are calculated based on small random sample (e.g., 100 examinee responses per pretest item). Once item latencies of all pretest items are available, pretest items are selected based on a latency difference index which equals the difference between 1) the interim pretest latency for the pretest items administered thus far and 2) a target value of the interim pretest latency specified for the current position. If the interim pretest latency is much less than the target, then a pretest item with large item latency is selected for the next pretest item. If the interim pretest latency is much greater than the target, then a pretest item with small item latency is selected for the next item. If the interim pretest latency is close to the target, then a pretest item is randomly selected. The initial item latencies of the pretest items are updated through the simulation as the sample size per pretest item increases (e.g., every additional 100 more responses per pretest item).

The standard deviation of the pretest latencies will be examined to determine if the proposed item selection method provides more comparability in the pretest latency across examinees. We also will examine whether the proposed latency control pretest item selection algorithm and the random pretest item selection are equivalent in terms of IRT parameter recovery and pretest item usage. The mean bias error (MBE) and the root mean squared error (RMSE) will be used to evaluate the stability of the estimated IRT parameters of the pretest items.

CLOUDCAT: Implementing CAT as a Web-Service

Haniza Yon (Mimos Berhad), Rense Lange (Integrated Knowledge Systems), Nur Ainshah (Mimos Berhad), Abid Altaf (Mimos Berhad) and Norsyahida Abd Kadir (Mimos Berhad)

Abstract:

Most CAT systems combine two basic functions: selecting the next question for a test-taker, and subsequently presenting this question and obtaining the test-taker’s answer. We are developing “CLOUDCAT” as a CAT solution based on a “software as a service” (SaaS) model – also known as “on demand software” – in which CAT functionality is entirely delivered by dedicated (e.g., cloud-based) servers. Clients will delegate all CAT related computations by sending requests to a server, and receiving answers, using standard web protocols (e.g., TCP, HTTP). This approach does not impose any software standards, and the service interacts identically with programs written in, say, JAVA, C, PHP, or Python. As a result, users can continue using their own item delivery and scoring systems, provided they adopt a small set of commands intrinsic to CLOUDCAT.

CLOUDCAT uses a socket-based approach, which implements a Rasch CAT server in Python V2.7 capable of accepting requests in TCP. CLOUDCAT has specialized commands to define items, specify item parameters and tests and subtests, update test-takers’ parameters, and define startup and termination conditions. This command set was designed to mesh optimally with the design and implementation of typical item administration systems. Additional functionality is planned for CLOUDCAT, and indeed is already being implemented in some respects. CLOUDCAT’s single most important command is “observation”, shortened to “o.” This command has five fields that contain all the information needed to update the test-taker’s position on the latent variable. The server reports this position, and determines the next item to be administered.

To date, research has focused on optimizing CLOUDCAT’s overall performance. It was found, for instance, that “o” requests take less than 0.01 seconds, even when more than 10,000 users are active simultaneously. However, the overhead incurred from the use of TCP-based sockets is a crucial question. Simulations are currently planned to address this issue, and the results will be presented.

Development of an Adaptive Integrity Test

Haniza Yon (Mimos Berhad), Norsyahida Abd Kadir (Mimos Berhad), Nur Ayu Johar (Mimos Berhad) and Nur Ainshah Abdul Mutalib (Mimos Berhad)

Abstract:

Computerised adaptive testing (CAT) is a powerful technology-driven method in which a computer programme administers test questions according to a dynamic algorithm that continuously estimates the ability level of the test taker and chooses successive questions accordingly. An important consideration in development of any CAT is the size of the item bank that will be needed to estimate test-takers’ ability with a satisfactory level of precision. This study examines the issue of item bank size and related parameters in the context of an adaptive integrity test that is currently being developed. Integrity is widely held to be a crucially important attribute, affecting numerous aspects of individuals’ behaviour both in and outside the workplace The test is intended to provide necessary information regarding a person’s level of integrity.

Monte Carlo simulation studies were carried out using the software package CATSim in order to explore combinations of item bank size and other parameters such as test length, item exposure, scoring algorithm and termination criterion that would be sufficient to determine candidate scores with particular levels of precision. Once an appropriate precision threshold for the test had been selected, the results from the Monte Carlo studies were used to determine how many additional items needed to be developed and added to the item bank in order to meet that threshold. The required new items were then developed by experienced item writers. All items were pilot tested, and then calibrated using a rating scale model. Post-hoc simulation studies incorporating real data from the pilot tests were also carried out in order to better predict how the CAT would perform with real candidates in the future. Although development of the CAT has been a lengthy process involving many challenges, live administration is expected to begin early in 2016.

Comparing CATS and the Block Review Method in Providing Review Options in CAT

Zhongmin Cui, Chunyan Liu, Yong He & Hanwei Chen

Abstract:

The research in the literature has shown that, if answer changes are allowed on a test, the final estimate of an examinee's ability could be more accurate because of reduced anxiety level and the opportunity to fix mistakes (Papanastasiou & Reckase, 2007; Wise, 1996). The practice of item review in computerized adaptive testing (CAT), however, is hindered by the potential danger of test-taking strategies (e.g., Kingsbury, 1996; Wainer, 1993) that exploit a typical adaptive algorithm. Examinees can either "trick" the system to administer easy items and make answer changes during review or revise wrong answers using the clue from difficulty changes. To provide examinees with review opportunities and minimize the impact of test-taking strategies, researchers have proposed to put some restrictions on reviewable CAT (e.g., Han, 2013; Stocking, 1997).

One promising approach is the block review method proposed by Stocking (1997). In this method, examinees can go back and review answers within a block of items. Once the answers for a block are submitted, examinees will move to the next block and cannot go back. Recently, Cui, Liu, He, and Chen (2015) proposed a new approach, referred to as Computerized Adaptive Testing with Salt (CATS), to implement CAT with no restriction on item review and answer changes. In their procedure, a mini test form is assembled before a typical CAT algorithm starts and test items from the mini test form spread out in the whole test in a way unknown to examinees. Because items from the mini test form do not depend on examinees' abilities, as Cui et al. found in their study, CATS is robust to the aforementioned test-taking strategies while providing examinees unrestricted opportunities to review and make answer changes to any test item at any time before submitting the whole test.

Both the block review method and the CATS method seem to provide a solution for reviewable CAT. It is unclear, however, that which method is better. On one hand, the CATS method seems to be more flexible than the block review method because of the elimination of restrictions. On the other hand, the block review method seems to be more efficient than the CATS method because only adaptive items are administered in the block review method. To address this question, we will conduct a simulation study to compare the CATS method and the block review method under various conditions. We will evaluate in terms of (1) the robustness to the test-taking strategies, and (2) the efficiency of CAT when no test-taking strategies are used. Bias, standard error, and root mean square error of ability estimates will be computed. The full details of the simulation procedure and results of the simulation study will be included in the final paper. We expect this study will provide guidelines for practitioners in administering reviewable CAT and provide both accurate ability estimates and scores, and a testing environment more conductive for examinees to demonstrate their real level of achievements.

Development of Multi Stage Tests based on Teacher Ratings

Stéphanie Berger

Abstract:

In a multi stage test (MST), test takers are adaptively assigned to test modules of varying difficulty levels depending on their results in previous test modules (e.g., Yan, von Davier & Lewis, 2014). Through optimizing the fit between the item difficulty and the test taker's ability, a MST can provide more reliable results over a broader ability range compared to a linear test. However, the development of a MST requires previous knowledge about the item difficulty in order to assemble the items to test modules of varying difficulty and to define the appropriate routing rules between the test modules of consecutive stages. Information about the item difficulty is usually collected by expensive and time-consuming linear pretests. Though, it is more and more difficult to reach students and teachers for participating in pretesting studies due to tight school timetables. Furthermore, pretesting under low stake conditions can lead to a biased item calibration if students' motivation and effort is low.

In spring 2015, a new set of standardized computer-based MST has been introduced in Northwestern Switzerland in order to measure students' ability in secondary school (i.e., grade eight) in different school subjects (i.e., German, English, French, and Mathematics). Each standardized MST consists of four stages including test modules of three difficulty levels. The practical circumstances did not allow to thoroughly pretest the items prior to the test assembly. Instead, secondary school teachers were asked to rate the difficulty levels of the developed test items. These ratings served - beside curriculum related content information - as the basis for constructing the different modules.

In our presentation, we will describe the applied MST design as well as the heuristics that were used to determine the routing rules in more detail. In addition, we will discuss the advantages of constructing a MST based on teacher ratings and elaborate on the challenges that are related to the absence of pretest data for the test assembly, the item calibration and the reliability of the test results. In order to validate the teacher ratings and the related assumptions that formed the basis for developing the MST design, we will present selected results from the administration of the tests to a sample of *N* = 8'001 secondary school students.

The design and development of a web-based adaptive Raven's-like automatic test generator

Isaac Thimbleby, Cardiff University

Abstract:

Presented here is an automatic item and test generator for matrix based intelligence testing.

Raven's Standard Progressive Matrices (SPM) is historically the most commonly used fluid intelligence test. Each of SPM's 60 test items consists of a two dimensional grid of elements governed by an overarching pattern (the matrix), with one element missing. Participants are asked to identify the correct missing element from a selection of possible answers.

Developing an automatic item generator for Raven's-like questions has a number of advantages:

1) Item generators are particularly useful in that they generate high numbers of test items, which is required for adaptive tests.
2) They can enable frequent testing, through extremely large numbers of test items generated.
3) They can be used as a tool to help improve our understanding of Matrix based tests.

A Raven's-like test generator capable of generating around 10^{300} unique test items was designed and developed. This number of test items is sufficiently large as to be unlimited for all practical purposes.

This approach was based largely on a quantitative analysis of Raven's SPM and the matrix component of Cognito to provide insight into some of the necessary design components, including heuristic based automatic difficulty assessment.

Further matrix features have been identified and refined through an on going iterative development process. Critical features identified include the ability of the tester to control test structure and characteristics in order to test hypotheses on test construction and completion; for example, hypotheses regarding the effect of a certain rule on difficulty. This would be particularly informative for the development of adaptive matrix tests.

Identifying Item Enemies in a CAT Pool Using a Chi-Square Goodness-of-Fit Test

Brian D Bontempo (Mountain Measurement, Inc.), Steve Wise (NWEA), Gage Kingsbury, Ron Houser

Abstract:

Many Computerized Adaptive Tests depend on large pools of IRT calibrated items which are assumed to be, amongst other things, statistically independent. In reality, operational CAT item pools may contain item enemies, items which are highly correlated that contain similar content. Detecting item enemies in a dichotomous multiple-choice question CAT is difficult because the assumptions required of traditional item correlation statistics are not met greatly reducing the validity of their use. This study posed a methodology for detecting item enemies and tested this method on an operational CAT.

The method used to detect item enemies used the final ability estimate obtained from a complete test, the anchored item parameters associated with the operational item pool, and the correct/incorrect results for each item response to conduct a simple chi-square goodness-of-fit test. A theoretical and an observed 2X2 contingency table were created for every pair of items. The theoretical (expected) table contained the estimates of the probability of a correct response to both items (1,1), the first item only (1,0), the second item only (0,1), and neither item (0,0) based on the final ability estimates and the item parameters. The observed table was created using the empirical dichotomous item results. The deviations between the observed and expected values were used to calculate the chi-square.

The data from a large-scale, variable-length, dichotomous Rasch model CAT used for licensure in the United States were used to evaluate this method. The complete data from a single operational item pool's use was extracted for analysis. In total, the data from 1,472 items and 39,681 examinees were analyzed. Only pairs of items that had been administered to at least 100 examinees were used in the analyses. In addition to calculating the chi-square described above, a tetrachoric correlation coefficient was calculated as a point of comparison.

The results of these analyses are scheduled for completion on August 1, 2015. This results of the analyses will include the number and percentage of items that were identified as item enemies using the tetrachoric correlation and the chi-square method. In addition, a 2X2 contingency table will be created which will quantify the similarities and differences between the two methods. The characteristics of pairs of items identified as enemies will be described including whether or not the items were mapped to the same topic of the test blueprint, the difference in the difficulty of the pair of items, and the sample size of the pair.

In summary, this study explained a simple chi-square method for identifying items enemies used on a computerized adaptive test, described the outcomes of using the method on a real operational CAT item pool, and compared the results to those obtained using tetrachoric correlations. The chi-square method promises to provide an easy to use, readily scalable method for identifying pairs of items that violate the assumption of statistical independence.

Estimating Reliability in Linear on the Fly Test (LOFT) Designs

Tammy Trierweiler

Abstract:

Linear on the fly testing (LOFT) involves the assembly of a unique fixed-length test for each examinee based on pre-specified content and psychometric constraints (Folk & Smith, 2002). In the LOFT design, either classical test theory or item response

theory (IRT) can be used to generate randomly parallel test forms, and these forms may be developed before the administration window or even during testing (hence “on the fly”).

Although LOFTs are operationally applied in several educational and certification testing programs, there is no published literature to date (that the authors are aware of) that deals specifically with the estimation of reliability in this framework. Building off work by Lord and Novick (1968) and Cronbach (1951; 1963), this study aims to address this gap by presenting a modification of the standard formula for alpha that can be used in the LOFT framework.

The concept of reliability is based on the notion of consistency or precision of a measurement, and different means of computing reliability estimates are appropriate for different test assembly/delivery methods. For a fixed test assembled following classical test theory (CTT) construction methods, coefficient alpha (Cronbach, 1951) may be computed as a function of item and score variances and used as a lower bound estimate of reliability. In the LOFT design however, as different subsets of items are randomly selected (based on content and psychometric constraints) to create unique examinee test forms, response data become inherently “incomplete” as only some of the items available for use from the larger item pool are delivered to any candidate. In this situation, it is not appropriate to use the standard coefficient alpha as a means to estimate reliability.

To address the LOFT reliability estimation problem, a modified alpha reliability formula is proposed that is rooted in generic true score theory (Lord & Novick, 1968, Ch. 7). Simply stated, the idea is to estimate the reliability of a randomly selected LOFT form (generic), rather than the reliability of any particular form (specific).

Analytic and Monte Carlo simulations were used to study the proposed modification to alpha. Item parameter estimates taken from a calibrated 668-item bank were used to construct 1,000 nominally parallel 90-item LOFT forms. Each form was constructed using pre-specified content and psychometric targets in the LOFT framework. For each constructed form, examinee response data were generated under the Rasch model using a standard normal distribution for latent abilities. Based on the information used to create the simulation, a theoretical value for the generic reliability of the test was obtained. In addition, sample values for the modified alpha coefficient were computed, and their sampling distribution compared to the theoretical reliability. This process was carried out for several different distributions of abilities and for several different sample sizes. For comparison purposes, a corresponding fixed test form was created using one of the LOFT forms. It was “administered” to all simulated examinees, and the standard coefficient alpha was also computed for each sample.

Moving the New Zealand Progressive Achievement Tests into the Computer Adaptive Domain

Hilary Ferral (Senior Statistician, New Zealand Council for Educational Research)

Abstract:

Progressive Achievement Tests (PATs) in New Zealand are a long-standing set of standardised assessments for students in Year 4 to Year 10 (approximately 8 - 14 year olds). PATs were first published by the New Zealand Council for Educational Research (NZCER) in 1968. Since then the suite of assessments has undergone a variety of modifications, modernisations and additions.

The current generation of PATs includes assessments in Mathematics, Reading Comprehension, Reading Vocabulary, Listening, and Punctuation and Grammar. Rasch measurement scales are used to report achievement. A wide range of customisable online reporting is available including class result lists, individual student reports, school-wide analysis, and longitudinal reports.

Over the last 10 years NZCER has been developing an online platform from which PATs can be administered. This began with the storage of data from paper and pencil tests and the creation of a small range of online reports. Now all of the PAT assessments can be delivered online. This has been a popular move with schools who appreciate the efficiencies of collecting data directly, and having instant access to reporting and analysis.

The next PAT development is to build a CAT system for PAT Mathematics. At our disposal we have:

- a ready-made bank of 584 well established and reliably calibrated multi-choice items currently delivered in the static tests
- information about population distributions at different year levels
- a set of stable year level norms.

In addition, over 100 new items - some interactive - are being developed to augment the bank.

This paper describes a simulation for a PAT Mathematics CAT which investigates the effects of various scenarios, constraints, and requirements. Starting points, ending criteria, item selection, item exposure, and content control issues are discussed.

We aim to deliver assessments for Year 4 to Year 10 students that

- provide a positive experience for the test-taker
- adequately target the ability level and year level of the test-taker
- provide teachers with useful information about next learning steps
- satisfy subject matter experts and test developers that proper coverage of relevant parts of the curriculum has been achieved
- can generate person estimates with adequate precision
- do not over-expose items
- provide flexibility for administrators to specify starting points and test length

These attributes are idealistic, and some are in competition with others. How much pressure do the ideals place on the current item bank? Which of the attributes should have priority over another, and to what extent?

The simulation provides some insight into how the PAT CAT will behave under different conditions, what outcomes we can expect at a population level, and what the individual test-taker may experience. This in turn will help to establish an informed set of conditions under which the PAT Mathematics CAT will operate when it is rolled out in 2016.

A Practical Model for CAT Development

Nathan Thompson

Abstract:

Thompson and Weiss (2011) presented a 5-step model for developing computerized adaptive tests (CATs). This model will be presented and discussed, then applied to real-life examples. It is unique in that most CAT research focuses on developing new quantitative algorithms, while this presentation is instead intended to help researchers evaluate and select algorithms that are most appropriate for their needs. It is therefore ideal for practitioners that are familiar with the basics of item response theory and CAT, and wish to explore how they might apply these methodologies to improve their assessments.

Steps include:

1. Feasibility, applicability, and planning studies
2. Develop item bank content or utilize existing bank
3. Pretest and calibrate item bank
4. Determine specifications for final CAT
5. Publish live CAT.

Computer adaptive measurement of quality of life across cultures: results from IRT and CAT-simulation studies.

Gibbons, CJ.[1,2] Skevington, S.[1], and the WHOQOL group.

1. Manchester Centre for Health Psychology, University of Manchester, UK.
2. The Psychometrics Centre, Judge Business School, University of Cambridge, UK.

Aims

To develop an item bank suitable for computer adaptive administration of quality of life (QOL) measures using the WHOQOL-100 instrument across diverse cultures. Simulated computer adaptive tests were conducted to assess item bank performance. We developed a strategy for presenting and scoring anchored-item banks across different cultures for implementation within Concerto V4.

Methods

Data from the international WHOQOL-100 field trial were psychometrically assessed using a combination of Mokken and Rasch analyses. The original 100-item bank was divided into four domains, representing the structure of the WHOQOL-BREF instrument.

Differential item functioning (DIF) was assessed using ANOVA with post-hoc Tukey tests. Where DIF was evident between countries, a strategy of 'splitting' items was employed to allow item 'difficulties' to vary across diverse cultures.

Computer adaptive testing was simulated using the FIRESTAR application to assess item bank performance (number of items administered) at different standard error thresholds using maximum posterior weighted information estimation.

Results

Mokken analysis confirmed the suitability of a four-factor structure consisting of domains measuring Physical QoL, a Psychological QoL, Social QoL and Environmental QoL. However not all items within each domain scaled uniformly (Loevinger's Ho < .30) and 22 items were removed prior to Rasch analysis.

Rasch analysis further reduced 55 items across all scales. Following an iterative process of item removal, all resultant scales showed excellent fit to the Rasch model (p>0.01), were reliable (Cronbach's α > .85), unidimensional and free from local dependency.

Whilst the items of the WHOQOL-100 have been shown to be conceptually equivalent, tests of differential item functioning suggested that a number of items provided different measurement information across different cultures. A complex solution which allowed item 'difficulties' to vary by country was used to allow valid pooling of data.

Where standard errors was set to .55 (equivalent to Cronbach's α = .70) the four item banks could be successfully administered using an average of 4 items each.

Conclusions

The current research demonstrates that brief QoL item banks can support fundamental measurement and perform well in simulated CAT situations. Approaching the issues of DIF by country in the manner demonstrated here may help to strengthen cross-cultural research and improve understand of quality of life.

Optimal Greedy Item Selection for Constrained Tests

Daniel Bengs

Abstract:

Nonstatistical constraints arise naturally in many areas of adaptive testing. Several approaches for constrained item selection in adaptive tests have been proposed including shadow testing and maximum priority index (MPI). In contrast to the heuristic MPI,

shadow testing is based on integer programming and guarantees constraint compliance. While shadow testing is a versatile and powerful approach, it is conceptually involved and the required optimization step is generally NP-hard; shadow testing trades off modelling complete, constraint-compliant item sequences and combinatorial complexity. Hence, commercial software libraries such as CPLEX are often the method of choice for efficient optimization, rendering the licensing of CAT programs difficult.

We focus on modelling smaller collections of items that we call feasible subtests. Feasible subtests are item sets that can be combined with additional items to form constraint-compliant tests. We propose a greedy item selection algorithm that assembles constraint compliant tests item-wise as follows: Given the set of already administered items, include the most informative item under the restriction that the new set forms a feasible subtest. The algorithm terminates when a maximal feasible subtest has been assembled, that is, a constraint-compliant test of maximal length has been found. Each step of the greedy item selection only requires number of feasibility checks that is linear the number of the remaining test items. Thus, irrespectively of the problem at hand, the required computation time is negligible.

The applicability and optimality of the proposed strategy depends on the algebraic properties of the set system of feasible subtests which is induced by the actual constraints. We derive conditions under which the greedy method computes fixed length tests that are globally optimal and thus coincide with tests assembled by shadow testing. We show that these conditions are met for various types of constraints such as mutually exclusive items or balancing of content categories. However, we also address limitations of our approach. For instance, constraints involving quantitative variables, such as bounds on the sum of response times, can in general not be treated in our framework. Empirically, we compare our approach to shadow testing on a real world item pool with balanced content category constraints and give implementation details.

In summary, the proposed greedy item selection method allows to solve various classes of constrained item selection problems in a considerably more efficient way than standard approaches such as shadow testing while guaranteeing constraint compliance and optimality of item selection. Our result provides a new theoretical perspective on constrained item selection and is also of practical relevance for the development of efficient constraint management in CAT programs.

How IRT helps LinguaLeo to build accurate placement procedures

Dmitry Abbakumov[1], Olga Degtiareva[2], Igor Lyubimov[2]

[1] National Research University Higher School of Economics, Moscow, Russia

[2] LinguaLeo, Moscow, Russia

Abstract:

LinguaLeo is a freemium online platform offering an English language learning service for Russian, Brazilian Portuguese, and Turkish speakers (www.lingualeo.com). The number of users is 12M. The first defining feature of the platform is that it allows each user to choose real-life content that he/she likes. The collection includes more than 300K learning materials, including news, entertainment and business articles, TED Talks, popular songs, movie clips, stories, and jokes. The second defining feature is use the personalization. The process starts from a placement test to determine the language skill level of each user. After the test LinguaLeo tunes the learning program to the user to make learning English more effective.

The test is adaptive and based on the grammatical frame. The frame consists of five levels of difficulty. Experts developed 262 items, sorted them by difficulty and packed them into the tree of decisions. The first version of the test did not use psychometric approaches. After using the first version during a couple of months, the LinguaLeo team found out the problem of effectiveness of the adaptive engine.

The purpose of the IACAT 2015 presentation is twofold: (a) to present the research of the psychometric quality of the first version, and (b) to share the results of developing the second version of the test based on IRT.

Firstly, during the research the items were calibrated in IRT on 1M sample of real users. We discovered several problems: (a) the poor quality of the tree and the lack of usage for 100 items; (b) the misfit of difficulty of items for the levels determined by experts, for example, items from the easiest first level were distributed from -3 to 3 logits on the scale. Secondly, we found and studied the differential functioning of items for Russian, Brazilian Portuguese, and Turkish speakers. Finally, we proposed two types of algorithms for the second version, which are based on IRT and the mix of IRT with the tree of decisions. These algorithms were tested in post-hoc simulations and under the conditions of the platform. In sum, the mixed algorithm works much better than the first version for the placement purposes. This algorithm starts and does several steps with IRT estimation procedures, and after involves the tree of decisions for item selection.

All the results will be shown in the presentation.

Two Ways to Unleash CAT's True Potential

Thomas Garrard

Abstract:

Simply moving away from paper & pencil is no longer enough. As Computer-Based Testing matures, Computer Adaptive Testing is playing an increasingly important role because of its many benefits: immediate feedback for teachers, improved efficiency for test-taking, accuracy for test results, as well as stronger security for deployments.

There is no doubt CAT is in the spotlight and making headlines. Now the question has become: how can we increase adoption and unleash its full potential? One way to do it is to make it plug-and-play and truly portable.

In this presentation, we will:

- Discuss how to leverage QTI and PCI for authoring CAT item banks, making them truly portable. While QTI itself isn't quite CAT compatible, QTI items are, which means you can use them to compose a CAT test. Meanwhile, PCI allows you to create innovative CAT items such as simulations, to make tests more engaging.
- Introduce TAO, an Open Source assessment platform that readily supports CAT items, from simple to sophisticated. TAO offers a CAT API so that you can easily plug your CAT engines to any TAO installations worldwide, and take advantage of the open source platform for either rapid application development or actual deployments.

Transitioning from Linear Testing to Multistage Testing: a Case Study

Maaike M. van Groen, Angela J. Verschoor, Hiske M. Feenstra & Ronald Engelen

Cito, Arnhem, the Netherlands

Abstract:

One of the main advantages of adaptive testing is the tailoring of the tests to the ability of the students. This results in tests in which an independent decision is made to select the next item(s). The Dutch Final End of Primary School Test (Cito, 2014; CvTE, 2015) is currently administered in two linear, mainly paper-based, versions. One version is intended for administration to the 25 percent of least able students and a version is intended for the remaining 75 percent of students. The pupils' teachers decide which version will be administered to each student. Since the test is used to provide an independent advice about the appropriate level of secondary education, the Dutch Parliament has decided that the format of the test should be changed to adaptive testing. This way, the selection of the items will be independent from the teacher. Independence of the teacher was considered important because the teacher also provides an advice to the student.

Although, the transition from a linear, mainly paper-based, test to a computerized adaptive test might seem straightforward it comes with a lot of challenges. One of the main challenges is to obtain pretest data for all the items that are required for adaptive testing. Another challenge with computerized adaptive testing is the risk of exposure of the entire item pool. Both

challenges can be reduced by limiting the amount of adaptivity of the test. A multistage test requires a smaller item pool and test developers can relatively easily replace an exposed segment by a new segment. Further advantages of multistage testing over computerized adaptive testing include the increased possibility of content control, the lower precision of item parameters that is required before test administration, and the software for test administration can be less complicated. Therefore, it was decided to transition to multistage testing.

Even with a multistage test a lot of challenges have to be overcome. However, a lot of interesting research opportunities also arise. To mention just some research topics:

- How many stages are optimal?
- How many segments per stage are optimal?
- How do we select the items for the segments?
- What is the optimal difficulty of the segments?
- How can the routing be arranged?
- How do we have to set up the cut-off values for the routing?
- How do we have to organize the pretesting?
- What can we report back to the students?
- How do we transform the linear report to a report based on a multistage test?

Some of these research topics were investigated in the literature and other topics were investigated using simulation studies. In the investigations special consideration was taken to the circumstances, limitations, and features of the current application. Data from the current version of the End of Primary School Test were used as input for the simulations.

A Successful Experiment of Computer Adaptive Test of Abilities in the Graduate Employability Development Domain

Dr. V. Natarajan, MeritTrac Services Pvt. Ltd., Rajeev Menon, MeritTrac Services Pvt. Ltd.

Abstract:

Higher education is at cross roads. On one hand, there has been a phenomenal and uncontrolled expansion of institutions, be they central or state, private or public, deemed-to-be or specialized, all of them produce a huge volume of educated manpower. Job opportunities are also increasing quite aggressively; yet many young graduates finding themselves not employed and under-employed. One of the real reasons for this is that there are more than a million jobs for which these graduates find themselves inadequately equipped with; particularly not so much of domain knowledge but on the cognitive abilities and skills that are increasingly insisted upon by all employers. MeritTrac has been focusing on these abilities test (Analytical Ability, Verbal Ability, Quantitative Ability, Attention to Details, General Awareness and Communication Skills)-often and inarguably considered as a good predictor of job success-and interpersonal skills as vital for any entry level job role. MeritTrac as a company felt that Computer Adaptive Testing to assess Graduate Employability Development Index in each of these abilities periodically, thrice during 3-5 years of graduation period (with a gap of at least one year between the tests) will release strengths and weaknesses of the test takers in relation to these abilities and skills. They can then work on the identified gaps, thereby helping them to see improvement and finally in the last exam, they will be given a probability index of meeting the requirements of a particular job role from the perspective of cognitive abilities. The adaptive tests in all these abilities have been created by R&D unit of MeritTrac and made available through a specialized platform capable of storing an item pool of 400 items in each of these abilities and using the 3 Parameter Model, calibrate these items including ICCs and Information Curves and creating a Fixed Length Adaptive Test of 12 items each and each candidate allowed to take all these tests and given the benefit of their True

Scores, Final Ability and GMAT like Scaled Scores with a feedback on the Strengths and Weaknesses of these abilities and their areas and suggesting remedial preparation to be made before taking the next test. For this, based on our cumulative experience and data responses of several items and several sets and for several test groups in the recruitment tests for corporate, government, public or private, a standard matrix has been compiled giving 4 levels of scores; top score ever, 27% of test taker group at the end of this group and the lowest end score, and 27% of lower ability group top score. This standard matrix will go on getting updated over the years. The experiment is so successful to start with that big corporates are very eager to implement this, given the obvious advantage of better predictability for their selection process and limit their bespoke recruitment tests to all those declared eligible by this Computer Adaptive Test series. The program can help them hire the same number of people more accurately with aspirants just 20% of the original.

An Exploratory Study of Starting a CAT with a Non-Scaled Item Pool

Deborah J. Harris, Chunyan Liu and Troy Chen, ACT , Inc.

Abstract:

To start a CAT, one piece of groundwork is, among others, to have a scaled item pool. However, implementing this requirement might sometimes face challenges for some testing organizations. For example, some Asian countries administer a new test each administration without linking and release all items after the test administration. In the US, data use is becoming a bigger issue, so a company might have items but not access to legacy data to calibrate them if they want to start a CAT.

The purpose of study is to explore the feasibility of starting a CAT when a scaled item pool is not available but the items are. A simulation study will be conducted to evaluate how different thetas are with a non-scaled pool versus a scaled one under CAT. The results of the study will provide some preliminary information on the impact due to using a substandard item pool when starting a CAT.

Item pool:

To construct a non-scaled pool for this study, 10 math forms each having 60 multiple choice items will be separately calibrated using real operational data under the 3-parameter logistic model. To mimic item parameters of released or compromised items, some item parameter estimates will be manipulated to make such items become easier and less discriminating. These 10 forms have same content specifications but no common items. Through a linking design based on special study data, the item parameter estimates of the 600 items can be placed on the same scale to constitute a scaled pool.

CAT design:

In this simulation study, a fixed-length CAT will be employed. Three test lengths of 60, 45 and 30 items are considered. For each test length, the distribution of content specifications of items administered will follow that of the original form (i.e., any of the 10 forms).

Data generation and analysis:

A random sample of 5,000 thetas from a realistic distribution will be employed to generate item responses under the CAT design using the non-scaled pool. Based on the simulated responses, theta estimates will be computed using the non-scaled and scaled item parameters and then compared. Several statistics including classification consistency and RMSE by different ability groups and overall will be assessed and reported.

Evaluating the Comparability of CAT Tests across Test Delivery Platforms

Agnieszka Walczak, Cambridge English Language Assessment, Ardeshir Geranpayeh, Cambridge English Language Assessment

Abstract:

Considerable amount of research has been concerned with mode effects, exploring the question whether the mode of test administration (computer-based or paper-based) may affect candidate performance in a test and candidate motivation and behaviour during a test. However, little is known whether any effects may arise which are related to computer platforms on which computer-based tests are administered. This article explores the question of comparability of the same test delivered via two different computer platforms. It employs live candidate data from a computer-adaptive placement test which assessed English language knowledge and which was administered to candidates via two different computer platforms. The article explores whether there are any differences with regard to candidate performance in the test administered via two different platforms and whether there are any differences with regard to how much time candidates spend on the test, each test section and on each question. In order to answer these questions, the article will employ propensity score matching. This proposal addresses a question that flows from operational work of an exam provider and has both theoretical and operational consequences for test construction.

Multistage Testing with Routing Based on Performance on Different Test Sections

Kyung (Chris) T. Han, Fanmin Guo, and Eileen Talento-Miller

Abstract:

Multistage testing (MST)—often viewed as a specialized version of computerized adaptive testing (CAT)—offers distinct advantages for both the test developer and the test taker over typical CAT formats. For example, MST offers test developers more control over test construction in terms of possible test forms and content combinations. At the same time, it enables test takers to move back and forth among test questions within each test section. A substantial tradeoff with MST compared with traditional CAT, however, is the reduced adaptability of the test, especially when a test has a minimal number of stages. In a typical MST administration, for example, the first stage usually is a routing stage in which all test takers see test items with the same average difficulty level. Assuming a test has three test sections, each measuring different traits, and each section consists of two stages with the first stage being a routing stage, then only three of the six stages (i.e., about a half of the test) would be adaptively administered—the other half would be of linear test administration. If multiple test sections measure different but moderately or highly correlated traits, then a score estimate for one section might be adequate for adaptively selecting item modules for following sections without needing to administer routing stages repeatedly for each section.

This presentation introduces a real-world case of test development with MST. The test consists of three sections, each measuring different cognitive skill sets that are moderately correlated with one another. Findings from a series of simulations with various conditions that show different correlational relationships among test sections will provide test developers with useful guidelines for determining how much improvement in measurement efficiency can be achieved under various MST situations by applying the proposed approach.

The development of a Computerized Adaptive Test with online calibration

Angela Verschoor,Cito, Stéphanie Berger, University of Zürich

Abstract:

Four cantons in northwestern Switzerland initiated the development of two assessment instruments: a set of compulsory standardized tests and a related online item bank for formative assessments. The use of computerized adaptive testing (CAT) enables students to evaluate themselves, and to get feedback about their current competences.

Until now, the development of large item banks needed for CAT involves extensive pretesting, data collection and calibration, before the test can go live. This is usually a complicated and expensive task in the development of an operational CAT.

In this paper we propose a new method to develop a CAT, specifically suitable for low stakes environments such as formative tests. The CAT will be built on the basis of an item response theory (IRT) calibrated item bank. Specifically, the Rasch model will be used. As one of the goals of the formative adaptive tests is to yield an estimation of the candidate's performance on the IRT scale underlying the compulsory tests, linking both scales is of utmost importance.

Question is, how we can maintain synchronization of both IRT scales during the life span of the item pool, while the online calibration will provide flexibility in item pool maintenance while keeping development costs low? First in a relatively small scale pretest using a multi matrix design a linking is established between both IRT scales using a number of strategically placed anchor items. Next to this an online calibration method will gradually add new items. The proposed online calibration consists of three phases, using different estimation methods In the first phase, an update according to the application of an Elo algorithm is used. This phase ends when a more extensive procedure yields more reliable calibration results. The second phase consists of a calibration with the Joint Maximum Likelihood (JML) method, which has the advantage over the Elo method of producing estimations with known, and generally much lower, error than the Elo method. Disadvantages of JML, however, are the fact that it has a biased estimator when used with CAT data, and that estimation errors tend to reduce only slowly with growing numbers of observations. Simulation studies have shown that JML is suitable up to approximately 100 observations per item, after which the effects of bias prevents further convergence of parameter estimations. Above 100 observations per item, the third phase will start, comprising an offline Marginal Maximum Likelihood (MML) estimation method, fixing all parameters of items that have attracted enough observations. At that time, new uncalibrated items can be added to the active pools.

In September 2015, the first item pools for German reading and Mathematics will be opened, using 40 to 50 anchor items per subject, plus approximately 400 uncalibrated items. Gradually more uncalibrated items will be added to the active pools, until the entire set of 1500 – 2000 items are calibrated using MML.

Skinning a CAT in the Real World: How An Educational Authority Transitioned a Low Stakes Assessment to CAT Using Concerto

Andrew Kyngdon

Abstract:

In Australia, the cost and reporting structures underlying traditional pencil and paper based testing programs are coming under increasing pressure. The key stakeholders of educational authorities, notably teachers, parents and schools, are demanding that assessment results be delivered within shorter timeframes. Persistent inquiries from stakeholders concerning online assessment are also being received. Interestingly, this is happening against a backdrop of considerable stakeholder apprehension concerning moves to put Australia's national assessments online; actions which have generated sustained, but at times negative, media coverage. In 2014, the Board of Studies, Teaching and Educational Standards commenced an investigation into transitioning the Record of School Achievement Literacy and Numeracy Tests from computer based assessments to adaptive ones. A simulation study, conducted by an independent organisation, established this as a feasible endeavour. Investigations conducted during the formulation of the business case suggested that the pricing models of various "off the shelf" solutions may not be congruent with the current financial environments in which statutory authorities must operate. Moreover, the development of institutional capability in CAT was identified as a key, strategic goal. Consequently, it was decided that the open source Concerto platform be used. A research trial conducted in 2014 yielded results that were consistent with the simulation study. End user acceptance was also positive. It was therefore judged that the Concerto platform was fit for purpose; and a decision was made to formally

deploy the assessments as “on demand” CATs on May 21 2015. This was successfully achieved. The most attractive features of CAT to key stakeholders, these being students, teachers and, in particular, the media, were the capacity for tailored testing, the immediate feedback to examinees at the end of the test, the fast production of formal credentials, the ability for examinees to sit the tests whenever they wished and the reputation of the University of Cambridge.

Stratification strategies for CD-CAT based on the linear/binary search

Chanjin Zheng

Abstract:

The restrictive progressive method (RP) and the restrictive threshold method (RT) are the only two methods developed specifically to address the item exposure control in CD-CAT. As information-based methods, they are computationally intensive. The current study proposes a stratification method to address the item exposure issue. The major advantage of the stratification method is much simpler and computationally much less intensive. But the extension is nontrivial since cognitive diagnostic models are drastically distinct from IRT models and there is no straightforward counterpart for the b matching step in CD-CAT. This difficulty can be overcome by using the linear and binary search algorithms in computer science. Searching problem has been intensively studied in computer science. Item selection in CAT can be considered as a searching problem (finding the best item for next administration) and this new perspective can generate some new idea easily.

Methods:

Algorithms. A general framework for the stratification method for CD-CAT consists of two parts: partitioning of the item bank according to the item discrimination index and selecting items by pattern matching method. Rupp provides an excellent summary of the item discrimination indices for DCMs which takes care of the first part. The pattern matching is not straightforward and two versions of pattern matching will be proposed from the linear/binary search perspective.

The linear pattern matching is straightforward. Assuming the current posterior distribution of the cognitive profiles is accurate, the ideal candidate item must be matched with the cognitive pattern with the largest probability.

The binary matching pattern is that the ideal candidate item must have such a mastery profile that can split the current posterior distribution in half with respect to the probability mass. The splitting process is very similar to the calculation of the eta in the DINA model. The item with the smallest value will be chosen for next administration.

Design. Two studies carried out to evaluate the two strategies. Study 1 aims to investigate the performance of the linear and binary pattern matching. In order to make the execution of linear pattern matching possible, the Q matrix was generated by randomly selecting from the distinct patterns and thus the number of items with every pattern was approximately equal. An item bank consisting 480 items of the DINA model with 4 attributes was used in the study. The item parameters gj and sj were both generated from U(0.05, 0.25). 2000 examinees were simulated in the way that all distinct patterns are assumed to be equally likely.

Study 2 aims to investigate the binary stratification strategy’s performance under more realistic situation in which the linear pattern matching is infeasible.

Evaluation criteria. 6 item selection methods were evaluated in estimation accuracy and item exposure balance (consult the table for the details).

Results:

The studies have been completed, but only the result for Study 1 is presented here due to the space constraint. The estimation accuracy and the measure of exposure balance of each method are reported in the table. With regard to the item exposure control, it is not surprising that the pure KL information-based method generates the largest chi-square value. That for the stratification strategy with binary pattern-matching is within the proximity of those of RT and RP, but the chi-square value of the linear search is still significantly larger. Number of overused and underused items can provide more information about the item bank usage. The linear search does not perform well in this aspect either and there is only some minor reduction in the number of overused items and the underused items is still as many as 259. RT, RP and binary search all did a good job in improving the item usage, the underused items is as few as about 35.

Relevance:

The item exposure rate issue leads to the test security problem and underuse of the item bank. Although cognitive diagnostic testing is often low-stakes, this cannot preclude the concerns for test security exclusively. The other issue is related to the prohibitively high cost of item development for cognitive diagnostic testing. Thus, item exposure rate also raises a practical question in the CD-CAT. The proposed binary stratification strategy is a much simpler alternative to the information-based methods.

Empirical comparison of scoring rules at early stages of CAT

David Magis

Abstract:

Usual scoring rules in CATs include maximum likelihood (ML), weighted likelihood (WL) and Bayesian approaches. However, at early stages of adaptive testing, only a few item responses are available so the amount of information is very limited and in addition constant patterns (i.e. only correct or only incorrect responses) are often observed, yielding ML scoring intractable. Specific scoring rules (such as fixed- or variable stepsize adjustments) were developed for that purpose. However recent research highlighted that both Bayesian and WL scoring rules may provide finite values even with small sets of items.

The purpose of this presentation is twofold: (a) to make a quick review of available scoring rules at early stages of CAT, and (b) to present empirical results from a simulation study that compares those scoring rules. More precisely, three scoring scenarios will be investigated: stepsize adjustment followed by ML, Bayes or WL followed by ML, and constant scoring rule throughout the CAT. These methods will be compared by means of simulated item banks and under various CAT scenarios for next item selection and stopping rules. Empirical results will be presented and practical guidelines for early stage scoring will be outlined.

An Application of Computerized Adaptive Testing in Audiological Assessment

Wayne M. Garrison and Joseph H. Bochner, National Technical Institute for the Deaf, Rochester Institute of Technology

Abstract:

Less than 15% of adults in the USA over age 70 receive hearing screening and less than 20% receive any form of treatment. Only about 25% of individuals who could benefit from a hearing aid actually use one. Reasons vary, but affordability and accessibility are major barriers to intervention and treatment. Unfortunately, hearing healthcare services are not covered by many health insurance plans and are unaffordable and inaccessible to many adults.

While the current model of hearing healthcare, predicated on clinic-based audiological assessment and intervention, will always remain the gold standard, new delivery approaches that translate research on scalable assessments and interventions are needed to expand delivery of hearing healthcare to underserved adults who do not benefit from traditional and current practices. To this end, accessible and affordable assessment and intervention protocols having demonstrated validity and reliability are needed.

In this presentation, we describe the development of a CAT protocol for hearing screening (NSRT: Bochner, Garrison & Doherty, 2015; Garrison & Bochner, 2015), focused on the measurement of speech recognition ability which, in combination with additional information provided by respondents (i.e., age and an indicator variable reflecting perceived hearing status) provide

the basis for the estimation of listeners' hearing thresholds (i.e., hearing sensitivities expressed in dB) across the frequencies 0.5 to 8 kHz. Hearing thresholds constitute the primary data used to program hearing aids.

The NSRT™ test materials are sentence-length utterances containing phonetic contrasts, primarily minimal pairs. The NSRT™ test protocol (accessible at https://apps.ntid.rit.edu/NSRT/) uses a paired comparison discrimination task in which a standard sentence is paired with two comparison sentences. Examinees must indicate if comparison sentences (delivered as auditory stimuli) are the same or different from a standard sentence (printed on the screen). Responses are scored correct/incorrect.

The NSRT™ was developed using the Rasch measurement model. The current item bank is the product of administrations of separate test forms (linked by common items) to two samples of respondents (n = 123). Respondents' performance on CAT versions of the test was found to be statistically equivalent to performance on common item test forms (p < .05). OLS linear regression analyses were then used to examine the relationship of NSRT performance (together with age and self-report measure of perceived hearing status) to hearing threshold measurements obtained in a clinical setting. The resulting regression models (developed from CAT administrations of the NSRT in quiet, and +5 dB SNR noise) revealed approximately 75% shared variance between hearing thresholds measured in the clinical setting and those predicted from data yielded by an interactive NSRT session. The presentation will focus on analytic/statistical details of hearing threshold prediction accuracy, notably at the casewise level.

Older adults are living longer and working longer. Their communication needs are greater than ever, and their proficiency and comfort level with new technology has never been higher. This presentation focuses on an application of CAT outside of education/psychology considered by some as a high priority in public health. Study findings have significant, direct applications in the delivery of hearing healthcare services.

An Improved Online Item Calibration Method for Multidimensional Computerized Adaptive Testing

Ping Chen, Beijing Normal University, Chun Wang, University of Minnesota, Jihong Xu, National Research Institute for Family Planning

Abstract:

Calibration of new items on-the-fly has been an important topic in both psychological and educational measurement. Multidimensional-Method A (i.e., M-Method A) has been proposed as an effective online calibration method for multidimensional computerized adaptive testing (MCAT) (Chen & Xin, 2013; Chen & Wang, 2014). However, M-Method A has a theoretical limitation (i.e., treating the estimated ability vectors as true values and ignoring the measurement errors in ability vector estimation that might be carried forward to item calibration) in the calibration process of new items, thus this method might yield erroneous item calibration when ability vector estimates contain non-ignorable measurement errors. To improve the performance of M-Method A, this paper proposes an improved MCAT online calibration method, namely, the full functional MLE-M-Method A (i.e., FFMLE-M-Method A). This new method combines full functional MLE (Stefanski & Carroll, 1985) with M-Method A with a goal to correct for the error in ability vector estimation that might adversely affect the item calibration precision. Moreover, two correction schemes are proposed when implementing the new method, depending upon whether the individual dispersion matrix (or error covariance matrix) or the average dispersion matrix is employed to correct for the measurement errors inherent in ability vector estimates. A simulation study was conducted to evaluate whether FFMLE-M-Method A can improve the calibration precision as opposed to the original M-Method A method under two levels of sample sizes (N = 3000 and 1500) and two loading structures (within-item design and between-item design). The results showed that the new method yielded more accurate item parameter estimates than the original M-Method A in almost all experimental conditions.

Application of the Adaptive Measurement of Change to Measuring Achievement Growth in Reading and Mathematics

Chaitali Phadke, David J. Weiss, Theodore Christ

The present paper analyzed intra-individual psychometric change pertaining to Math and Reading skills. Examinees were measured using the fixed-length (30-item) Adaptive Measurement of Change (AMC) method. Measurements were taken at three or two time points. A total of 65,633 K-12 students completed CATs for their Math ability at Time 1 (early in the school year), 53,684 students were measured at Time 2 (midway through the school) & Time 1, and 38,997 students were measured at Time 3 (the end of the school year), Time 2 & Time 1. Reading CATs were administered to 179,225 students at Time 1, 153,946 students at Time 2 & Time 1, and 101,683 students Time 3, Time 2 & Time 1.

Detecting Significance of Change

When the examinee is measured repeatedly over t time points, with respect to his/her ability (q) level, his/her q may be estimated at each of the t time points. The hypothesis of no change can be stated as $H_0 : q_{t_2} = q_{t_1}$ versus an alternative of $H_1 : q_{t_2} {}^{1} q_{t_1}$. The significance of change can be determined by testing the set of hypotheses. In the current analysis, a $Z-$test, proposed (in another context) by Guo & Drasgow (2010) was used with $a = 0.05$ at an intra-individual level to determine if significant change had occurred within each individual examinee. This $Z-$test has the following form:

$$Z = \frac{\hat{q}_2 - \hat{q}_1}{\sqrt{SE_2^2 + SE_1^2}}$$

where, $\hat{q}_2$ = CAT estimated ability at Time 2, $\hat{q}_1$ = CAT estimated ability at Time 1, SE_2 = standard error of measurement associated with $\hat{q}_2$, SE_1 = standard error of measurement associated with $\hat{q}_1$. With the property of local independence under item response theory (IRT), $\hat{q}_2$ and $\hat{q}_1$ can be assumed to be independent and the Z statistic to follow a standard normal distribution under the H_0.

Three different Z statistics were computed for the students who were measured on all three occasions, pertaining to the three difference scores ($\hat{q}_2 - \hat{q}_1$, $\hat{q}_3 - \hat{q}_2$ and $\hat{q}_3 - \hat{q}_1$).

Results:

Table 1 presents the proportion of students who demonstrated significant change for Math and Reading.

Table 1: Proportion of students with significant change between two ocassions for Math and Reading datasets

	Math	Reading
Time 1 – Time 2	0.30	0.26
Time 2 – Time 3	0.25	0.15
Time 1 – Time 3	0.60	0.43

As Table 1 shows, the proportion of students showing a significant amount of change for Math ranged from 0.25 to 0.60, with the maximum change observed between Time 1 and Time 3. A similar pattern was observed for Reading in the range of 0.15 to 0.43. Figures 1 and 2 show scatterplots of change scores plotted as a function of $\hat{q}_1$ for Math and Reading, with data points identified as significant or insignificant change. More cases were detected as significant as the difference scores moved away from 0. Minimum significant change was detected at the difference score of 0.39 for Math and at 0.28 for Reading. The difference scores correlated near zero with $\hat{q}_1$ for Math ($r = -0.12,$ 0.08 & -0.07) and were moderately correlated for Reading ($r = -0.44,$ -0.20 & -0.59). Additional analyses for students measured across three time points and who had two or more significant changes indicated that the vast majority of change was nonlinear, falling into three different patterns. The results support the use of the AMC procedure for the detection of psychometrically significant change with elementary school Math and Reading data.

Multidimensional IRT versus Mean Adjustment in Estimating Cognitive Ability Scores

Darrin Grelle

Abstract:

Since the online employment testing boom began, client demand for more information in less testing time has greatly increased. Clients want more data points upon which to make hiring decisions, while at the same time reducing the testing burden on their candidates. This demand requires that psychometricians find ways to either glean more information out of each question or to find ways to estimate more accurate scores with fewer questions. This paper seeks to do the latter. It is common for job candidates, typically recent college graduates, to take between three and five different cognitive ability assessments. A single cognitive ability test, even an adaptive one, might take the candidate as long as 30 minutes. If the candidate is completing a full battery of tests, this candidate might sit for a testing session that lasts for up to 2.5 hours. The goal of the present research was to evaluate three different methods of administering cognitive content in an effort to determine the most parsimonious way of estimating cognitive ability.

In this study, 200 items from three different cognitive ability tests (deductive reasoning, inductive reasoning, and numerical ability) were administered to volunteers from the graduate population. Each candidate saw 12 items from each facet and the item administration algorithm rotated evenly through the three facets. The test was timed, but not speeded. Each of the individual facet tests is typically 18 items. After collecting data for 12,000 individuals, three different IRT models were estimated: a correlated factors MIRT model, the bifactor model, and three unidimensional IRT models (item parameters were estimated for each facet separately). Fit statistics indicated that the bifactor model was a poor fit, but the other two models were of equivalent fit. Theta for each of the three facets was estimated using the two well-fitting models. They were highly correlated. After using the score adjustment method described by Wainer (2001), the adjusted unidimensional estimates were almost perfectly correlated with the MIRT theta estimates.

The results of this study indicate that scores on three different cognitive ability tests can be accurately estimated using both MIRT and unidimensional IRT with a simple post-hoc adjustment and a 33% reduction in testing time. These results indicate that more sophisticated methods are not always necessary to achieve psychometric goals. The next phase of this study will evaluate the extent to which test length can be reduced using these methods while still providing an accurate estimate of ability.

Item Selection Method Based on Attribute Mastery Probability for CD-CAT

Xiaofeng Yu

Abstract:

Cognitive diagnostic computerized adaptive testing (CD-CAT) is a popular mode of online testing of cognitive diagnostic assessment (CDA). The key to a CD-CAT system is the item selection strategies. Some of the popular strategies are developed based on some information measures, such as Kullback–Leibler information (KL), Shannon entropy (SHE), to select items in CD-CAT. Typically, during CD-CAT, these familiar methods would use a cutoff point to transform the attribute mastery probabilities' provisional value to binary values, this cutoff point method may lead to a larger deviation. A method that can take advantage of the probabilistic information with regard to attributes may offer a better alternative.

Based on this consideration, this paper proposed two item selection strategies based on the provisional value of the attribute mastery probabilities, as follows: (1) The first one is called as SAMP1 (the strategy based on attribute mastery probability), and it can lead to maximum difference of the sum of attribute mastery probabilities. The SAMP2 makes use of more information currently available of examinees. (2) The second one considers the fact that not all the patterns are equally likely, but overlooks the fact that the distances between different patterns and the current estimation are not all of equal importance.

To investigate the performance of the proposed two methods, three simulation studies were carried out, one was the fixed length of CD-CAT which imitates a 20-item-length CD-CAT, and the second was the variant length CD-CAT which the test termination rules are the maximum length reaching 20 or the posterior pattern probability reaching 0.8, and the last was the short length CD-CAT that the test length are 4, 6, 8 or 10. Variant item selection strategies were taken into consideration in these studies, including KL, SHE, PWKL, SAMP1 and SAMP2. The results were compared in terms of pattern or attribute classification correct rate, item average exposure ratio, item maximum exposure ratio, item minimum exposure ratio, average test length, unused item number, number of items with exposure ratio over 20%, test overlap ratio.

The simulation results indicate that: (1) the comprehensive performance of SAMP1 and SAMP2 are better than other mentioned strategies in fixed and variant length CD-CAT; (2) SAMP1 and SAMP2 can retain a good measurement accuracy, and also improve the utilization ratio of item pool. It also should be noted that the mentioned item selection strategies still need to be improved in terms of the utilization ratio of item pool because all the strategies mentioned before choose less than half of the items in the simulation pool.

The first stage items selection methods of CD-MST

Chunlei Gao, Zhaosheng Luo, Chanjin Zheng, Xiaofeng Yu, Peida Zhan

Abstract:

Cognitive Diagnostic Multi-stage Testing (CD-MST) is a new testing mode which combines Multi-stage Testing (MST) with Cognitive Diagnostic Assessment (CDA). It is adaptive at the stage level, and conformity to our customs, it also allows administers to check the pre-assembled test forms, and the examinees to review and revise answers.

The initial stage items selection method is one of the most important elements for CD-MST, because CD-MST is not adaptive within the initial stage, and it has less adaptive frequency, the estimation of the first stage would influence the final recovery rate greatly. It is presented seven initial stage items selection methods. These methods can also extend to CD-CAT. They are Random Method, Item Selection Strategy, R* Matrix Method, Basic Concepts for Item Discrimination Index (BCIDI), Information-Based Item Discrimination Index (IBIDI), BCIDI with R* Matrix Index (BCIDIR*), IBIDI with R* Matrix Index (IBIDIR*). Random Method means that select first stage items from item bank randomly, this method does not consider item parameters and attributes, it is easy and fast. Item Selection Strategy is first to assign a knowledge state to examinees randomly, and then select

the initial stage items according to item selection indices (such as PWKL). R* Matrix Method is to select initial items from R-matrix. Rupp (2010) pointed out that in DINA model, based on CTT concept, the discrimination index of an item is 1-s-g, which means the probability of correct response to an item for respondent who has mastered all the attributes measured by an item minus the respondent who has not mastered the attributes. So here we use 1-s-g as the selection method, and called it BCIDI. Rupp (2010) also used CDI as the discrimination index, CDI is a stable index, and it does not depend on the knowledge state of examinee, so it can has higher accurate under the initial stage of the testing. Here we use CDI as another first stage items selection method, and called it IBIDI. We also combined BCIDI and IBIDI with R* matrix, called BCIDIR* and IBIDIR*.

A simulation study was presented here to judge the efficiency and accuracy of above methods. The number of examinees is 6000, the number of attributes is 5, item bank has 1240 items, test length is 15, it is said that the quality of item bank can influence the recovery rate, so here we used item parameter as a contributory factor.

The results presented that the Random Method was the poorest one, and the BCIDIR* was the best one. After the first stage, Item Selection Strategy was the lowest, it recovered quickly, after the whole testing, only a little poorer than discrimination methods. The discrimination methods and R* Matrix Methods were good methods but was influenced by the item quality, especially the item bank variance. Because the variance of item bank can determine the quality of selected items.

Effects of errors in item parameter estimates on recovery of theta estimates in CAT

Alper Şahin – Cankaya University and David J. Weiss – University of Minnesota

One of the advantages of computerized adaptive tests (CAT) over paper-and-pencil tests (PPTs) is CAT's capacity to match the items to the estimated ability of the examinees. However, it is generally assumed that in order to obtain optimum item and examinee ability match, the item parameters estimated will be estimated accurately so that the ability parameters estimated during the test will be accurate. Because all estimates entail some degree of error, this process cannot be deemed as error free. This means that the error in item parameter estimation before CAT application might affect the accuracy of CAT examinee scores and because item selection in CAT is based on information, which is derived from the estimated item parameters. For example, if the discrimination parameter is high, item information increases so that an item has more potential to be selected as the next item in a CAT. If the discrimination parameter estimate is incorrect due to estimation error, it could cause CAT to use an inappropriate item as the next item. Hambleton and Jones (1994) studied the effects of estimation error in item parameter estimates on item information functions. They found that the error in the discrimination parameter estimates inflated item information functions 25% to 40% when a sample of 400 examinees was used.

This study was designed to investigate how CAT ability (θ) recovery was affected by errors in the item parameter estimates, which typically result from calibration sample size and numbers of items. For this purpose, CAT item bank size and the sample size used to calibrate item parameters were manipulated. A data set of 10,000 examinees and 500 items was simulated. The samples (150, 250, 350, 500, 750, 1,000, 2,000, 3,000, 5,000) and the items for item banks (500, 300, 200) were drawn from this full data set. Item parameters were estimated in each sample. Then, θs of 10,000 examinees in the full data set were estimated using these item parameters through post-hoc simulations. Correlations, root mean squared difference (RMSD), average signed difference (ASD), and average absolute difference (AAD) values were calculated for both item and θ parameters. One-way ANOVA and regression analyses were conducted using these error statistics for ability and item parameters (*a* and *b*).

Results indicated that sample size (after 250) used to estimate item parameters and bank size had no significant effect on the examinee θ estimates in CAT. The estimation error in the *a* parameters had a more significant effect on the errors in θ estimates when the error was calculated as RMSD and AAD. The error for the *b* parameter had more effect on the error in ability estimates when ASD was used to calculate errors. Errors in CAT θ estimates are driven mostly from the error in *a* parameter estimates;

however, such errors in *a* parameter estimates did not curtail θ estimation accuracy. Thus, the results suggest that large samples are not required to develop useful CAT item banks if the primary use of the item parameters is to estimate θ using CAT.

Response Formats and Trait Estimation Efficiency in Computerized Adaptive Testing

Yin Lin, University of Kent&CEB, Anna Brown, University of Kent

Abstract:

Self-report personality assessments are dominated by single-stimulus (SS) response formats, where respondents provide rating responses to each item on a pre-defined scale (e.g., the Likert scale). Although easy-to-use, this response format is susceptible to response styles such as acquiescence and leniency (Van Herk, Poortinga, & Verhallen, 2004), and is sensitive to response distortions such as faking (Birkeland et al., 2006). The forced-choice (FC) response format has been proposed as a solution against such response biases. Forced-choice response formats ask respondents to consider multiple items simultaneously and rank them according to their own preferences. The ranking process removes all biases acting uniformly on all items (Cheung & Chan, 2002), provides resistance to score inflations associated with impression management and faking (Jackson, Wroblewski & Ashton, 2000; Christiansen, Burns & Montgomery, 2005) and improves differentiation between scales being measured (Bartram, 2007). Moreover, when analysed using an appropriate Item Response Theory (IRT) model, normative scores can be recovered from forced-choice responses (Brown & Maydeu-Olivares, 2011, 2013). Forced-choice response formats are thus becoming more popular especially in high-stake assessment situations, where the accuracy and fairness of measurement is at risk in the presence of response biases and distortions.

While Computerized Adaptive Testing (CAT) techniques for single-stimulus items are relatively mature, forced-choice CAT remains an area needing further research. Brown (2012) demonstrated that CAT successfully improves trait estimation efficiency over random item selection for a forced-choice Big Five personality assessment. However, the efficiency of forced-choice CAT has not been benchmarked against single-stimulus CAT. One may suspect that polytomous single-stimulus ratings would naturally give more information for trait estimation than binary forced-choice pairwise comparisons, thus making forced-choice CAT less efficient. But is this actually the case?

Extending from the Brown (2012) study, this present study explores CAT trait estimation efficiency using a multidimensional forced-choice pair format and a four-point single-stimulus rating scale. Assuming construct equivalence across the two response formats, a simulation study will be conducted using the MAT package in R (Choi & King, 2014). The simulations will utilise an item bank covering 10 competency scales measured by ≥25 items each, with item parameters derived from single-stimulus trial data (N=4113). One set of item parameters (SS1) is estimated using unidimensional IRT models for each scale (RMSEA≤0.03 and CFI≥0.95 for all scales). Another set of item parameters (SS2) will be estimated using a bifactor model, controlling for a general method factor that may be affecting single-stimulus data. The forced-choice item bank will be constructed by creating all possible multidimensional pairs from the single-stimulus item bank, with no content balancing. CAT simulations will then be conducted using A-optimality with Fisher information matrix as suggested by Brown (2012). Accuracy and efficiency of CAT convergence will then be compared across conditions (SS1 vs SS2) and response formats (SS vs FC), by looking at trait estimation bias and reliability against varying form lengths. The implications of findings on forced-choice CAT will be discussed.

Improvement and evaluation of a small-scale ESP CAT

Tetsuo Kimura (Niigata Seiryo University) and Yukie Koyama (Nagoya Institute of Technology)

Abstract:

Kimura, Ohnishi & Nagaoka (2012) implemented a Rash-based computer adaptive test (CAT) module to a major open source learning management system, Moodle, based on a BASIC program named UCAT developed by Linacre (1987). Kimura & Koyama (2015) constructed an in-house small-scale English for Specific Purposes (ESP) CAT with an item bank of lexical items, which were developed through the analysis of relevant corpora, specifically corpora of science magazines and academic engineering journals. After a couple of CAT administrations with the item bank consisted of 180 lexical items, the authors found a shortage of lower difficulty items. In order to improve the CAT, relatively easier items were created through corpora analysis of science textbooks in English-speaking countries and another 61 lexical items were added to the item bank. This paper will report the results of CATs before and after the improvement of the item banks and discuss advantages and disadvantages of a small-scale ESP CAT.

A Variable Length CAT for Ranking Data in Forced Choice Assessments

Shungwon Ro, IBM; Chia-Wen Chen, Wen-Chung Wang and Xue-Lan Qiu, The Hong Kong Institute of Education

Abstract:

Use of forced choice assessments to measure multidimensional personality traits has gained popularity in the workforce setting. Attempts have been recently made to resolve a well-known issue of ipsativity in forced choice assessments using new item response theory (IRT) models. We introduce a newly developed forced choice computerized adaptive testing (FCCAT) program using a generalized logit model, one of new family IRT models for ranking data (Qiu, Wang & Ro, 2015). Initial item pool building began with constructing two forced choice forms, based on a nonequivalent anchor test design. Each test form, including 66 triad items with an anchor set (48% common), was constructed with 132 statements (11 per trait, covering 12 traits), and pretested to a sample of 1,600 participants selected considering balanced demographic conditions such as country, gender, ethnicity, education level and employment status.

Model feasibility was checked via a simulation study. A triad item consists of three statements, each measures different trait/dimension. On each item, six ranking patterns were identified and treated as categories of a polytomous item. The results showed good parameter recovery in both the statement utility (location parameter) and person parameters. With all statements calibrated using concurrent calibration, an initial item pool was finalized with a selected set of statements for each trait. This produces an item bank of 26,620 triad items for an item selection during a CAT session, which may cause a computational burden and a processing time issue (i.e., very long latency).

A variable length FCCAT was designed to allow an item bank to be created on the fly (not a full-blown but sub/partial bank) during the CAT session to address the challenge. The Fisher information was used for item selection with a sub-pool selection/construction procedure to reduce the computational burden and save processing time. In addition, two exposure control strategies were adopted to prevent over-exposure of particular statements in each CAT session. Traits were estimated using the Bayesian maximum a posteriori procedure. As for the termination rules, a fixed standard error of measurement (SEM) cutoff (< 0.45) and a maximum test length (sixty six triad items, which is equivalent to a static version of FC assessment measuring 12 traits) were applied.

Results of simulation studies indicate that equ-precise measures were obtained with roughly a half of the triad items, which are needed by the random item selection (i.e., non-adaptive) procedure. Using two exposure control strategies, within-person exposure of each statement during each CAT session was well controlled at a pre-specified level. The item processing time, from displaying an item to selecting next item, was controlled under 0.8 seconds, particularly with one of three proposed sub-pool selection strategies. The results are promising for the application of the new IRT model and FCCAT program toward numerous non-cognitive assessments. The implication of the simulation studies will be discussed.

Heuristic Constraint Management Methods in Multidimensional Adaptive Testing

Sebastian Born

Abstract:

Multidimensional adaptive testing (MAT) is an approach to measure multiple traits simultaneously. Particularly when the dimensions are high correlated, MAT turned out to be more efficient than using multiple unidimensional adaptive tests. Nevertheless, until now there are hardly any operational assessments in which MAT is used. Even in the field of large-scale assessments, in which complex competences are measured, operational applications are still rare. One explanation for the limited use might be the management of test specifications which is more complex in MAT than in CAT.

Especially in standardized testing programs, it is inevitable that different test forms are comparable regarding a predefined set of test specifications (e.g. number of items per dimension, total testing time, answer key). Constraint management methods (CMM) such as the shadow test approach (STA), the weighted penalty model (WPM) and the maximum priority index (MPI) provide a solution for this challenging task by modifying the item selection algorithm of an adaptive test to consider the criteria of statistic optimality and the required test specifications simultaneously. The STA has proven to manage test specification successfully in both, unidimensional and multidimensional adaptive testing. However the implementation of STA requires considerable knowledge in linear programming and the use of expensive solver software if the set of constraints is complex. Against this background, heuristic CMM are an important alternative for practitioners having a high potential to foster the operational use of MAT. Because the size of test assembly problems is a crucial factor, it is important to known how the different CMM perform for different numbers of constraints.

Therefore, the present study focuses on the effectiveness of two promising heuristic CMM in MAT for a varying number of imposed constraints. A full factorial design with two independent variables (IVs) is used. The first IV *Constraint Management* (none, WPM, MPI) compares the item selection based solely on the criterion of statistical optimality with the two heuristic CMM WPM and MPI. With the second IV *Constraints* (3, 13, 23, 33, 43, 53), the total number of constraints is varied. The evaluation criteria for the performance of the two CMM were the measurement precision and the extent to which the imposed constraints were fulfilled.

The study reveals that the MPI and the WPM are capable to handle complex sets of constraints in MAT while fulfilling all constraints without any violation. In contrast to using no CMM, the observed percentages of tests with violation are quite high. Furthermore, it can be stated that the measurement precision for both, WPM and MPI, decreases with an increasing number of imposed constraints. However, the loss in measurement precision is different for the WPM and the MPI. While the MPI seems to be more suitable for assessment situations with a low to medium number of constraints, the WPM should be preferred for large number of constraints. Based on the results it is recommended to consider the size of a test assembly problem before choosing a CMM.

ABSTRACTS FOR SYMPOSIA

New Developments in CAT by Graduate Students at UIUC

Shiyu Wang

Abstract:

This coordinated session introduces some new research on CAT developed by graduate students at the University of Illinois at Urbana-Champaign (UIUC). The five papers address different issues in two broad areas of current research, namely 1) using response time to improve various aspects of CAT and 2) exploring new item selection strategies in CD-CAT.

Study 1, **Online Calibration Strategies for a Joint Model of Item Responses and Response Times in CAT**, explores online calibration strategies that capitalize on an auxiliary variable in response times (RTs) to cope with the problem of data sparseness in online calibration**.** Additionally, it develops adaptive field-testing strategies under the D-optimality scheme by varying the level of information adoption from the RTs.

Study 2, **Maximizing All the Information: Using Response Times in CAT,** proposes the use of response times (RTs) with item responses in CAT to enhance item selection and ability estimation. Using van der Linden's (2007) hierarchical framework, information functions for item selection in CAT are developed. Information functions are derived following Ranger's (2013) approach.

Study 3, **Utilizing Response Time in Sequential Detection of Compromised Items in CAT**, proposes the incorporation of response times into a sequential procedure for detecting compromised items (Zhang & Lu, 2014) by simultaneously monitoring significant decreases in response times. Evaluating abnormal changes in both the frequency of correct responses and response times for items may provide additional insights into the nature and severity of the compromise.

Study 4, **Stratification strategies for CD-CAT based on linear/binary search**, proposes a stratification method to address the item exposure issue in cognitive diagnostic computerized adaptive testing (CD-CAT). The stratification method is conceptually much simpler and computationally much less intensive than the existing method. However, the extension is nontrivial since cognitive diagnostic models are drastically distinct from IRT models and no straightforward counterpart to the *b* matching step exists in CD-CAT. This difficulty can be overcome by using linear and binary search algorithms that are widely implemented in computer science for studying searching problems. Framing item selection in CD-CAT as an algorithmic searching problem inspires new research directions and approaches.

Study 5, **The Relationship between Q-Matrix Specification and Item Exposure Rate in CD-CAT,** re-expresses the maximum Kullback-Leibler (KL) information item selection criteria in CD-CAT as a combination of noise and Q-matrix loadings and examines the relationship between item exposure and Q-matrix specification through simulation studies. Additionally, it investigates the generalizability of results to a PWKL-based item selection in which a posterior weight is added for each attribute profile.

Some bothersome problems for operational CAT and some potential solutions

Gage Kingsbury, Brian Bontempo & Anthony Zara

Much of our CAT research is done in simulation, or in laboratory settings, where we control many of the aspects of the test setting. While this research is crucial to the improvement of CAT, it doesn't always tell us how our tests will operate in practice. This symposium addresses issues that are commonly encountered in the implemntation of an adaptive test that aren't easily simulated. Each of the speakers has encountered the issue that they address while putting an adaptive test into the field. They have been intimately, and sometimes painfully, involved in decisions concerning the issues they will discuss. Each speaker will briefly discuss the issue and its manifestation. They will then discuss the steps that were taken to try to deal with the problem, including the options that didn't work, as well as those that did. Following the presentations, the speakers will discuss these issues and other operational issues with the audience.

Tony Zara will begin the session with discussion of procedures that might be used for making decisions for candidates who run out of time. Regardless of the time limits we set for a testing session, some candidates will always still be working when time runs out. This is a critical issue for certification and licensure testing, because the decision made may allow the candidate to move forward in their profession. Tony will describe the identification of reasonable time limits, and will give examples of the consequences of using several different rules to judge those who run out of time.

Brian Bontempo will discuss the implementation of CAT in rapidly changing certification and licensure settings. As testing for certification and licensure spreads thorughout professional and technical fields, many agencies that would like to implement adaptive testing may be limited by small item pools, small candidate samples, and by rapidly changing test content. Brian will describe the difficult decisions to be made by agencies facing some or all of these challenges.

Gage Kingsbury will discuss the identification of idiosyncratic knowledge patterns for an adaptive test in education. While IRT gives us a strong base for creating scales and adaptive tests, students tend to ignore IRT when they are responding to an adaptive test. Each test taker will use their idiosyncratic knowledge patterns to respond to test questions. Gage will discuss procedures that might identify students with particular patterns of performance to help teachers make appropriate decisions.

Producing CAT scores for test takers who run out of time

Anthony R. Zara

In operational testing programs, all tests must practically balance a combination of testing time and number of items administered in order to meet the dual constraints of providing reasonable measurement accuracy for all test-takers and not incurring unreasonable test costs. A significant advantage of CAT programs is the efficiency of the measurement due to the optimization of information gained per item answered. CAT programs for licensure and certification needing to reach a known level of accuracy for decision-making will generally terminate under three conditions: (1) sufficient information is gained to make a confident decision; (2) maximum number of items reached; or (3) maximum amount of test time achieved. Conditions 1 and 2 allow for a graceful ending, with test takers reaching a decision that meets psychometric requirements for validity. Condition 3, however, allows for test takers to stop testing (by running out of time) before reaching the required level of measurement accuracy.

In licensure and certification test situations, it is not tenable for the test taker to complete the testing time available, yet not complete the exam and have the test administrator inform the test taker that he/she cannot receive a score because she/he has not taken enough test items to provide sufficient information to make a pass/fail decision. Every test candidate must be given a result for each test taken.

The key question is: how are we able to make valid pass/fail decisions for test takers who have not provided the level of information needed to make a “comfortable” decision. This presentation will describe for one operational licensure testing program the impact of candidates who run out of time (ROOT), conditions that lead to candidates running out of time and the steps taken to mitigate that impact.

The presentation next will discuss the options for managing ROOT results in operational programs, how the organization decided to implement the current rule and some information about its impact on licensure decisions.

The presentation will close with additional information related to time limits, fairness, and how the organization is able to communicate these complicated psychometric issues to stakeholders.

CAT for Rapidly Changing Content Domains

Brian D. Bontempo

Two challenges that most CAT programs face are keeping the content current and keeping the item parameter estimates current. These challenges are especially difficult for CAT licensure and certification programs which are faced with a rapidly changing content domain such as those in information technology or the medical sector. In medicine, protocols which are adopted and endorsed by international organizations often dictate practice. When these protocols change, practice changes instantaneously and the CAT program must adapt to these changes as quickly as possible.

In a more traditional CAT program, adapting to changes in the content or protocols would entail the following:

- identifying the outdated or erroneous items and either removing, retiring, or turning them off,
- writing some new items which reflect the new content or protocol to replace those that were removed,
- pretesting the new items,
- and calibrating & linking the new items to the scale.

Typically, this process takes over a year, which is way too long to keep medical practitioners up to speed with the latest protocols especially in emergency medicine.

At the same time as the items are being written and pretested, the education system is also adapting to the changes in the content. In the US, there are often over one thousand different educational programs for a single medical profession and even more for some of the basic information technology professions. The extent to which these programs are aware of the changes varies. Some actively keep up with changes while others ignore the changes. Most of these programs utilize learning resources such as text books and websites which need to re-written or revised to keep up with the content changes.

Given the fluidity of the changes in education, the extent to which candidates answering the newly written pretest items are exposed to the new content during their education and training varies dramatically. This makes it extremely difficult to obtain items that survive pretesting. Typically, the new items are way too hard and/or have very low point biserials correlation coefficients due to the erratic nature of the way in which the content is known across the population. Even if an item survives pretesting, it is likely that the empirical item difficulty estimate will not be stable and the item will be flagged for parameter drift soon after it is made an active item.

Another issue faced by CAT licensure and certification programs is when the eligibility requirements for an examination change. For example, CAT program may need to adapt their program when the duration of the required education and training is lengthened by a substantial amount. Faced with the opportunity to provide more education to candidates, educational programs will spend additional time on certain topics while ignoring others. The impact of these changes on the scale can be enormous and must be monitored.

The presentation presents ways in which programs can address these issues and identifies research opportunities. Some of the solutions may reflect actions taken by active programs while others are more theoretical in nature.

Can CAT give us knowledge about idiosyncratic performance?

G. Gage Kingsbury

The nature of educational testing is a constant struggle between less and more. The time that a teacher has with students is precious and testing can be seen as time taken away from instruction, so less testing is desirable. At the same time, if a test is given, educators would like to use the outcomes for as many things as possible, so more information is desirable. This tug-of-war has allowed educators to see adaptive testing as a very desirable tool, which allows short tests and accurate outcomes.

Unfortunately, educators often expand the use of adaptive test outcomes beyond the uses for which the test is intended. Recently, tests designed to categorize students as “basic”, “proficient”, or “advanced”,

have been put to use to measure individual growth and teacher effectiveness. This extension of the use of test scores is commonly not warranted by the test design. Unfortunately, the test design is hardly ever considered by the groups setting educational policy, or by the classroom teacher who needs to make educational decisions every day.

The current study examines a situation in which educators wanted to use an adaptive educational achievement test in mathematics to identify idiosyncratic learning patterns. While this use was beyond the scope of the original test design, the nature of an adaptive test allowed the investigation of the possibility. This presentation describes the nature of the request, and the actions that followed. (The presentation will also define the difference between a multidimensional latent space and an idiosyncratic learning pattern.)

First, the presentation describes the need that teachers had for more than just a single mathematics score. In a classroom, a single score in mathematics may be useful in assigning marks and in identifying students with special needs, but a teacher needs more. Specifically, information about goals within mathematics that are strengths and weaknesses for a student would be quite useful. This would allow a teacher to use flexible grouping of students, and would allow pinpointed catch up work for specific students.

Second, the presentation describes the initial response to the need, and some of the inadequacies in that response. Specifically, goal scores could be created for each student, given the content constraints used in the adaptive test, and so they were. Unfortunately, even with a reasonable number of goal areas (4 to 6), the standard error of the resulting goal scores were quite high, and tended to have only slight utility in the classroom. This doesn’t meet the educators’ needs but suggests opportunities for improving the testing design without substantial increase in test length.

Finally, the presentation describes the new adaptive test design that was proposed to allow information about the latent trait to be collected, as well as the desired information about idiosyncratic patterns of knowledge. To finish, the presentation details the political and logistic considerations that have kept this new design from being adopted.

CAT around The World

Organizer & Chair: Alina A. von Davier, Educational Testing Service,

Panelists: Tetsuo Kimura (Japan), John Barnard (Australia), Marié de Beer (South Africa), John Rust (UK), Theo Eggen (The Netherlands), Andreas Frey (Germany), Mariana Curie (Brazil) and Alina A. von Davier (USA)

Abstract:

This panel brings together testing experts from different parts of the Globe to share their experiences with adaptive systems. While making the world flatter (Friedman, 2005), we are also hoping to make it better by learning from each other and by respecting and preserving the local needs and policies. Given that the most salient use of an adaptive test is when the population of test takers is heterogeneous and that the testing populations do become more heterogeneous everywhere in the world primarily due to social mobility and migration of various reasons, a conversation about the usefulness of CATs, challenges in designing efficient algorithms, and debates about fairness and score reporting is necessary. In addition to having each presenter discuss the adaptive tests in a specific country or groups of countries, the panelists will also address questions about the newest technological advances implemented in their part of the world, the newest research results that were implemented in the specific adaptive tests, and the biggest challenges they encountered. The purpose of this panel is to provide a forum for the panelists and the audience to interact on best practices in operational practice of CATs.

Reference: Friedman, T. L. (2005). *The world is flat: A brief history of the twenty-first century*. New York: Farrar, Straus and Giroux.

CAT Developments in Japan

Tetsuo Kimura

Niigata Seiryo University, Japan

In Japan CAT has been applied in only a few large-scale educational assessments. One of the earliest practical uses of CAT in Japan is CASEC (computer assessment system for English communication), an English proficiency CAT for Japanese, which more than 100,000 people annually take. Another one is J-CAT (Japanese computerized adaptive test), a Japanese proficiency CAT for foreigners, which is mostly used for their placement in educational institutions in Japan. Meanwhile, several CBTs have been put into practice in different fields and some are in developmental stage. A few of them are seeking possibility to develop CATs in the near future.

CAT Down Under – Developments in Australia

John Barnard

EPEC, Australia

CAT is relatively new in Australia and the first substantial application thereof is in the assessment of overseas trained medical practitioners who wish to practice in Australia. Since candidates take the exam worldwide, the development of "parallel" non-CAT exams for different time zones increasingly became problematic and a viable solution was found in CAT. The transition from a paper-and-pencil through online to CAT will be briefly discussed together with issues dealt with. There are also rather early developments in educational assessments where CAT is being considered.

CAT in South Africa

Marié de Beer

M & M Initiatives, South Africa

The use of CAT in South Africa only started in the late 1980s but its development and general use has been slow on the uptake. Initial CAT development was aimed at student selection with the aim of providing quick yet accurate assessment of applicant students for selection purposes. A CAT version of two parallel versions of a standard scholastic aptitude test in the early 1990s was followed by a custom-made CAT for a dynamic (test-train-retest) measure of learning potential in the late 1990s which used two linked adaptive tests. More recent CAT developments include the initial steps toward CAT versions of two different personality measures. Projects in the pipeline also include the use of automated item generation in a CAT administration – using African art and artefacts inspired cognitive items. The use of CAT is still very limited in the South African (and African) contexts. Modern theoretical developments and freely available software programs and platforms are likely to have a positive impact on future CAT development and applications in the South African and Africa context.

Computerized Adaptive Testing in the United Kingdom

John Rust

The Psychometrics Centre

University of Cambridge, UK

In the UK we saw the first roll-out of a national Rasch-based national assessment of school performance in 1976, only for it to fall foul of the political landscape. Although 40 years have since elapsed the debate is still on-going. I will evaluate the objectors' arguments then and now, and also consider the role of CAT within the wider field of Computational Behavioural Science, where advances in machine intelligence are transforming the landscape.

Computerized Adaptive Testing in the German Speaking Countries

Andreas Frey

Friedrich Schiller University Jena, Germany

Even though several research groups are focusing on CAT-issues in the German speaking countries (Austria, Germany, Liechtenstein, Switzerland), larger operational applications of CAT are still rare. Only in the German speaking part of Switzerland adaptive tests are already periodically used in large-scale assessments of student achievement. In Germany, the Federal Employment Agency applies several tests for career counseling purposes. Besides that, several IRT-based adaptive tests had been developed in the German speaking countries and are sold by commercial test publishers. Additionally, several adaptive tests had been developed for the assessment of clinically relevant person attributes like depression, anxiety, stress, and obsessive compulsive disorder and are used in psychosomatic departments of hospitals. Recently, the development of adaptive tests was and is fostered by research initiatives funded by the German Federal Ministry of Education and Research. The developments include adaptive tests for the measurement of student achievement in reading, mathematics, and science, professional competences of medical assistants, and key variables in orthopedic rehabilitation (everyday functioning, depression, physical functioning impairments). In the presentation, I will give an overview of the CAT-related research activities, available adaptive tests, and operational CAT programs in the German speaking countries. I will then focus on the development of three web-based adaptive tests measuring student achievement in reading, mathematics, and science, and will briefly introduce the multidimensional adaptive testing environment (MATE) which is used for the delivery of the three tests. Finally, I will give an outlook on future developments in the area of CAT for the German speaking countries.

CAT in (Education in) The Netherlands

Theo Eggen

CITO- University of Twente

The Netherlands

Since the 90’s of the last century a limited number of CAT applications are operational in the Dutch educational system. Initially CAT was mainly used for summative assessments but nowadays most applications are used in formative assessment settings. In the presentation the multi segment adaptive testing approach used for development of CATs at Cito will be described. Furthermore a short overview of operational CATs in other fields than education will be given.

CAT developments in Brazil

Mariana Curi

ICMC – University of São Paulo

In this talk, I will present the research methods for CAT that have been studied in Brazil and the efforts that have been taking place for the implementation of some of them. Most of the CAT programs in Brazil have been developed to evaluate proficiency in different educational content area; some research and CAT programs are for psychological testing. And last but not least, some efforts have been focused on the construction of testing platforms that allow users to create their own CATs.

Adaptive Tests in the USA

Alina A. von Davier

Educational Testing Service

USA

In the USA the adaptive tests are used in many applications, from psychology, to licensure, to educational teaching and educational assessment. Moreover, there are different types of designs across these uses: some tests use an item-adaptive level algorithm (CAT), some use a module-level adaptive algorithm (MST), while the educational learning systems use learning algorithms from the machine learning and artificial intelligence that maximize the learning objectives, such as the ‘multi-armed bandit” algorithm. In educational assessment, we have seen a proliferation of the adaptive tests. Since there are many high-quality CAT practices in the USA, this presentation will be confined to K-12 CAT assessments and some of ETS‘ MSTs.

IACAT 2015 EARLY CAREER RESEARCHER AWARD WINNER

Optimal Design and Scoring for Adaptive Multi-Stage Testing: A Tree-Based Approach

Duanli Yan

Multi-stage testing (MST) is an algorithm-based approach to administering tests in stages with groups of items at each stage. As with CAT, the current applications of MST rely heavily on item response theory (IRT). Thus, these MSTs can have some of the limitations of CAT when IRT assumptions are violated, for example with multidimensional tests (Yan, Lewis and Stocking, 2004). This research introduced a tree-based MST algorithm which has no model assumptions. It also compared a range of potential MST designs using small sample calibrations (*N*=250) and their applications to a large real operational assessment.

Methodology

The tree-based MST provides efficient routing and scoring in constructing and analyzing MSTs without the aid of strong IRT models.

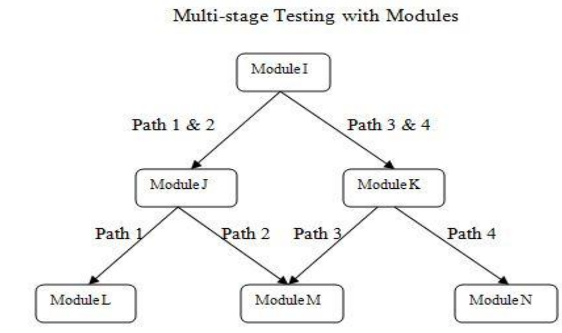


Figure 1. An Example of a Three Stage Multi-stage Testing Structure

Module scores. X_{11} is denoted module score on the first stage Module I; X_{21} and X_{22} are module scores for Module J ("easier") and K ("more difficult"); X_{31}, X_{32} and X_{33} are module scores for Modules L, M and N ("easiest", "intermediate", "most difficult"). All module scores are the number of correct responses to the items in the module. *Criterion score.* The total number correct score *Y* for a test consisting of all items in the pool.

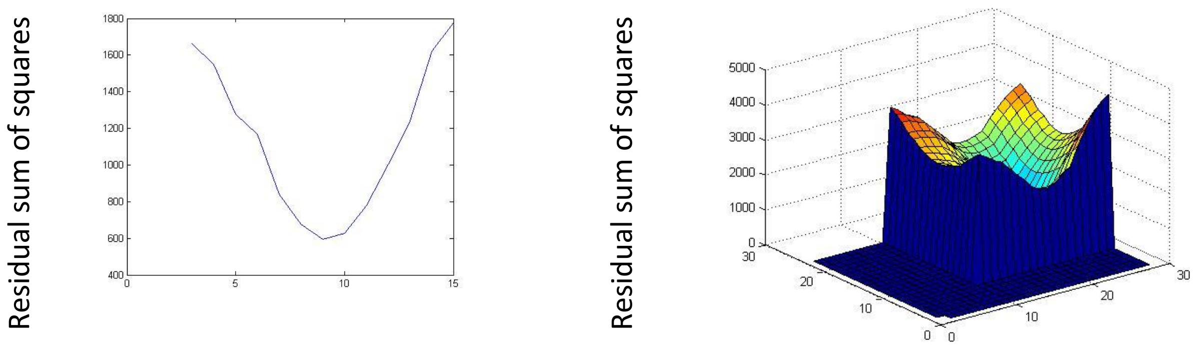
Cut scores. c_{11} is the cut score for the first stage Module I. If $X_{11} < c_{11}$, the test taker is administered the easier Module J; if $X_{11} \geq c_{11}$, he/she is administered the more difficult Module K. At the second stage, c_{21} and c_{22} are the cut scores for test takers who have completed Module J or K. If $X_{11} + X_{21} < c_{21}$, they are administered Module L; if $X_{11} + X_{21} \geq c_{21}$, or if $X_{11} + X_{22} < c_{22}$, they are administered Module M; if $X_{11} + X_{22} \geq c_{22}$, they are administered Module N. The goal is to predict the total score *Y* as well as possible, based on the paths and the administered module scores.

Finally, the total score *Y* is estimated $\left(\hat{Y}\right)$ by a set of four linear regressions of this criterion on the observed number-correct scores for all the modules at all three stages for each subsample.

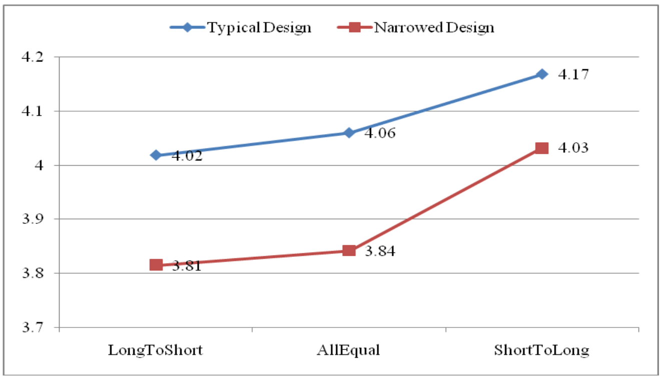
The results for several MST designs are evaluated in an application sample by two commonly used statistical measures: The proportion of criterion variance accounted for by the predicted score, and the root mean squared error.

Implications

This research data is from a real operational assessment rather than simulations, so the MST models, designs and the results are not artificial but have real world implications. The MST models were based on a random sample of only 250 actual test takers from the large data set. The small sample size is often an issue for 2-parameter and 3-parameter IRT-based model calibration. Thus, for real world situations when IRT-based MST model assumptions are not satisfied or only small samples are available, this tree-based MST can be an efficient alternative approach. This method and applications are included in the edited volume *Computerized Multistage Testing: Theory and Applications* by Yan, von Davier and Lewis (2014).



Stage 1 and 2 Cut scores



RMSE for Six Designs

ABSTRACTS FOR STUDENT GRANT WINNERS

A Partial Likelihood Method for Computerized Adaptive Testing to Allow for Response Revision

Shiyu Wang

The debate about whether to allow examinees review items and change answers in Computerized Adaptive Testing (CAT) has continued for nearly 30 years. The main concerns of allowing such options in CAT are about losing estimation efficiency and reducing test score validity by some test-gaming strategies. In this study, we propose a flexible CAT design that allows for response revision by using a partial likelihood method based on the nominal response model. In our design, we consider a multiple choice test situation where each item has $m \geq 3$ categories. Examinees are allowed to revise responses to any item at any time during the test. The only restriction is that for the same item, examinees can only revise it up to $m-2$ times. By using the nominal response model through a partial likelihood method, all the response data points, first attempt and revisions, are counted into the examinee's ability estimation on the fly. We study theoretically the proposed CAT design and showed that the ability estimation efficiency from such design is the same as or even better than that from the regular CAT design where response revision is not allowed.

Besides the theoretical results, we also investigate the performance of our proposed design in three test-taking scenarios, one in which examinees were simulated to make several careless mistakes, and the other two in which examinees were simulated to take two well-known cheating strategies. The results indicate that by using our proposed design, 1) serious examinees who made careless mistakes during the test can have chance to correct those errors, and this can result in smaller estimation biases and root mean square errors than when they are not allowed to go back to revise errors in a standard CAT design. Especially, comparing correcting all the errors at the end of the test, it's better for examinees correct those mistakes once they realize them during the test to be more accurately estimated. 2) Examinees who took the Wainer strategy were heavily penalized in their ability estimation and had a high risk to get a very low score. 3) Examinees who manipulated the Generalized Kingsbury strategy during the test gained nothing from such game-strategy, and their ability estimates were similar as those when they were not allowed to change answers in a standard CAT design.

Our proposed design is flexible in the sense that examinees are allowed to revise responses to any items freely during the test. This provides examinees a more relax test-taking environment and can help reduce their anxieties. Allowing for such options in a CAT can provide examinees chances to correct mistakes which may enhance test validity. To implement our proposed design in a multiple choice format test, test developers only need to calibrate item parameters from the nominal response model through a pretest, and develop a scoring algorithm by using the partial likelihood method.

Rules induction based method for the item selection in computer adaptive testing

Maria Rafalak

The aim of this work is the implementation of the rules induction method for the selection of items in the computer adaptive testing dedicated for the testee classification. Sequence of the responses for each testee may be treated as the training example for the discrete version of the rules induction algorithm (having limited and discrete values of item responses), such as AQ. Its purpose is to generate rules allowing for distinguishing between the selected trait/ability levels based on responses to particular items. The general scheme illustrating the idea is presented in Figure 1.

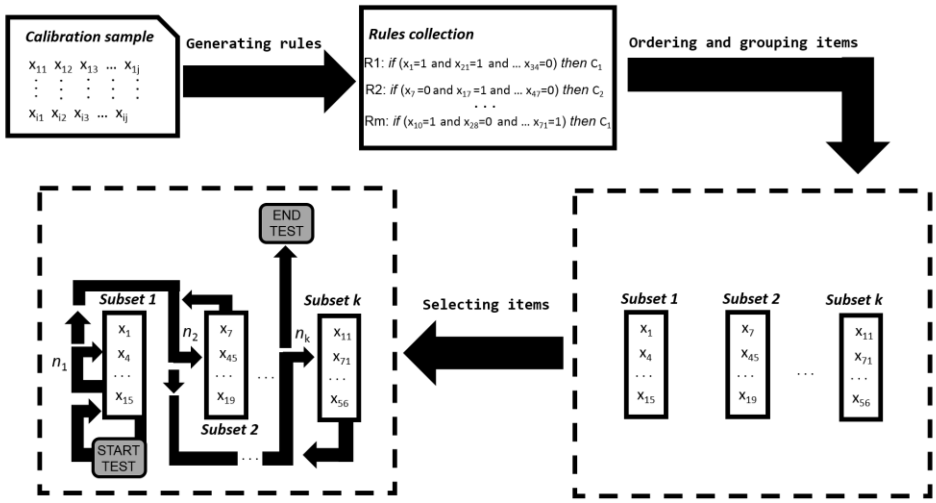

Calibration sample
x_{11} x_{12} x_{13} ... x_{1j}
x_{i1} x_{i2} x_{i3} ... x_{ij}
Generating rules
Rules collection
R1: *if* (x_1=1 and x_{21}=1 and ... x_{34}=0) *then* C_1
R2: *if* (x_7 =0 and x_{17} =1 and ... x_{47}=0) *then* C_2
. . .
Rm: *if* (x_{10}=1 and x_{28}=0 and ... x_{71}=1) *then* C_1
Ordering and grouping items
Subset 1
Subset 2
Subset k
x_1 x_4 ... x_{15}
x_7 x_{45} ... x_{19}
x_{11} x_{71} ... x_{56}
Selecting items
END TEST
START TEST
n_1
n_2
n_k

FIGURE 1. General scheme of the rules induction based CAT.

Like in most machine learning applications, the preparation of the training set requires large calibration sample. The rules premise part contains indexes of subsequent items, while the conclusions is the measured trait/ability level category (***C***). There are multiple different rules, using various items (***x***) allowing to distinguish between examples in the training set. The proposed algorithm generates the maximum number of rules (***R***) of the same quality which are then analyzed regarding the premises (the subsets of items allowing to distinguish between various categories). Quality of the rule reflects the number of examples (both real examples from calibration sample and hypothetical examples) covered by the rule. Rules that are fairly general are preferred in the generation process.

Generating many rules allows for ordering the items based on their frequency of occurrence in the rules collection. Popularity of the items reflects their discriminating power. Items present in many rules are more important for distinguishing examples from the training set.

After ordering, items are grouped into smaller subsets with the same or similar frequency of occurrences in the rules collection. The items from such subsets are randomly selected for administration. This will allow for uniform distribution of items among discriminating parameter and should facilitate the control of item-exposure rate. This approach is similar to the alpha-stratified adaptive testing method.

The first item in the test will be randomly selected from the subset of questions occurring relatively rarely in the rules collection. The next steps include drawing ***n*** items from the subsequent item subsets differing in item discriminating power. The ***n*** parameter can be either predefined by the test constructor or depend on the answers provided by the testee. As for the termination criterion, the fixed test length and minimization of classification decision variance are to be compared. Apart from classification decision final score can be estimated using standard IRT models.

During the conference I am planning to present the general overview of the proposed machine-learning-based approach to the computer adaptive testing together with the examples of its application. Computations will be executed both on simulated data and on the dataset from the Polish normalization and validation study of the Culture Fair Intelligence Test (CFT20-R) carried out by the Polish Psychological Tests Laboratory of the Polish Psychological Association in 2011.

Self-Adapted Testing as Formative Assessment: Effects of Feedback and Scoring on Engagement and Performance

Meirav Arieli-Attali, Advisor: David Budescu, Fordham University, New York

The merits of adaptive testing are not only in its psychometrics properties, but also in motivational aspects; adaptive testing has by design the potential to reduce some inherent motivational issues resulting from boredom (too easy tasks) or frustration (too difficult tasks). Most adaptive testing is conducted within the framework of Computerized Adaptive Test (CAT), where the computer program presents items selected by an algorithm. The focus of this study is Self-Adapted-Test (SAT), an adaptive test that allows TTs, instead of the computer program, to choose the difficulty level of the items. Like CAT, SAT relies on Item Response Theory models. It was first propose by Rocklin and his colleagues (e.g., Rocklin & O'Donnell, 1987), as an alternative to CAT, primarily as a way to reduce anxiety in high stakes tests. Most algorithms applied in CAT are such that maximize measurement (minimize measurement error), resulting in test takers being assigned items that roughly have 50% probability of answering them correctly. In SAT, test takers can choose items with higher probability of correct response (easier items), or lower probability (harder items), according to their motivational state, and their willing to put effort and challenge themselves. Studies investigating SAT have found that SAT holds desired psychometrics properties as well as is advantageous in reducing anxiety (e.g., Rocklin & O'Donnell, 1987; Vispoel, 1998; Wise, Plake, Johnson, & Roos, 1992).

This study examines variants of SAT in assessment framework that has no personal consequences to test takers, particularly interested in increasing *intrinsic* motivation via the design of the test. In particular, the study examines the potential benefits of SAT in the *process* of learning – as part of a *formative assessment* framework (Black, Harrison, Lee, Marshall, & Wiliam, 2003; Black & Wiliam, 1998a, 1998b), rather than just as an *outcome* of learning. Due to the SAT's unique characteristics, this form of testing permits exploring different ways of guiding TTs via the feedback opportunities, while observing effects on the self-regulated behavior (choices of difficulty selection). In an experimental design, the study manipulates feedback about total score (termed here "global feedback"), updated and presented after each question, to inform test takers continuously about their performance, and in that to allow them to adjust their choices. Four types of global feedback are compared in a between-subject experiment and with a control group (no global feedback): each type either provides external or internal incentive to put more effort (select harder items), or both. The study is currently in the stage of data collection and results are not obtained yet. The hypotheses are that TTs who are instructed to excel at the test (performance goal) will be particularly motivated by the global feedback that serves as external incentive. However, under a learning goal, TTs will be less sensitive to the external incentive, but rather react to a global feedback that is relative and reflect progress.

Latent Class Based Item Selection for CAT in Progress Tests

Nikky van Buuren, MSc

This research project has attempted to extent already existing methods for the construction of a computerized adaptive test(CAT), with as a goal to create an adaptive test which measures progress.

The challenge found in developing computerized adaptive progress tests(CAPT) with reliable and precise scores for the examinees is the underlying distribution of the abilities of the population. A CAT, usually has an item response theory(IRT) model applied to the data which can be used to construct an item bank to select items from. However, many items in progress data do not follow a form of the IRT models. This eliminates the possibility to calibrate an item bank using IRT and to create an item selection algorithm which uses the estimated ability of an examinee based on previous reponses to select the item providing maximum information on the examinees true ability.

The item bank for a progress test generally consists out of two different type of items. The first type of item would be a item in which the ability of the examinees gradually increases over time, which are named growth items. While for the second type of items there is a sudden increase in ability at a certain point in time, the so called jump items. The growth items can be associated with educational goals learned by experience over years, and jump items are taught at a set point of time in the curriculum of an examinee.

To be able to select from an item bank with these different type of items, a new method is proposed to measure progress in a CAT, which is based on a latent class model with 2 or 3 latent classes. The method used is an application of Cheng's proposal to use latent class models for CAT to perform cognitive diagnosis (2009). The estimated probabilities to answer correctly to items in a progress test when belonging to any of these classes are used to calculate Kullback Leibler(KL)-information for all the items. The KL- information values can then be used to select items and to construct an adaptive test. Simulations show that item-selection based on KL-information outperforms random selection of items in progress testing. Finally, a scoring method based on latent class probabilities in the CAT is proposed.

References:

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD CAT. *Psychometrika, 74(4)*, 619-632.

Weighted-Probability Based Classification for Computerized Classification Testing

Victoria Song

Classifying a student into one of several groups based on her test score remains a research problem in the areas of computerized classification testing (CCT). The Sequential Probability Ratio Test (SPRT) is well studied, initially proposed by Wald (1947) and further developed by Spray (1993) for classification in CCT. However the SPRT methods always results in an indifference region in decision space in which no classification decision can be made and requires more test items to reach a decision. We attempt to develop a weighted-probability method that always provides a classification on the student score. The method will improve classification accuracy and efficiency, especially for two-stage testing when stage 1 classification determines the groups and the type of stage 2 testing.

Assume that the test score of a student is given by

$$X = T + E \tag{1}$$

Where T is true score, E is a random error that account for the uncertainty of the test because of random factors affecting the test outcome. The effect of E could be measured by the test reliability:

$$R_x = \rho_{12} \tag{2}$$

Where ⍰$_{12}$ is the correlation if a student take a pair of parallel tests.

Two-category classification is to assign a student with a score X into one of two states (1 mastery, or 0 non-mastery). The goal o this study is to design a classification rule that is both accurate and fair to students under tests. The simplest rule is the hard-classification rule, in which a cut-off, X$_{cut}$, is derived. The student is classified as 1 if X is above X$_{cut}$, and 0 otherwise. The hard-classification rule is unfair to students of 1 if their test scores close to but slightly below the cutoff. In order to improve accuracy and the fairness, we introduce a probability function over a score region to classify students whose scores are close to the cut-off value.

The weighted-probability classification method is illustrated in the diagram with a piece-wise probability function for a two-category example. The same analogy applies for multi-category classifications. The test score is divided into 3 segments separated by two cut-off, X_{upper} and X_{lower}. If the score X is above X_{upper}, the student is classified as 1. If X is below X_{lower}, the student classified as 0. If X is between $X_{upper}$and X_{lower}, a new and variable cut-off value X_{cut} is randomly chosen within the score between $X_{upper}$and X_{lower} with a probability function, P(X), and the student is classified as state of 1 if the score X >X_{cut} or 0 if X <X_{cut}.

Selection of X_{upper} and X_{lower}, and P(X) are the focus of this study. To determine their optimal values, a simulation is used to generate student's true and test scores. Different probability functions are applied and tested. For each type of P(X), values of X_{upper} and X_{lower} are varied to affect the outcome of classification. ROC (receiver operational characteristics) is employed to measure the performance of different classification rules. This paper will present and discuss the results of this study.

Item selection criteria for Logistic Positive Exponent model-based Computerized Adaptive Testing

Thales Akira Matsumoto Ricarte, Mariana Curi, Alina Von Davier

The aim of Computerized Adaptive Tests (CAT) is to administer a personalized test, selecting items from a bank according to the examinee's abilities. The most common item selection procedure has been based in maximizing item information. Functions based on Fisher and Kullback-Leibler information are the most commonly used for item selection when Item Response Theory (IRT) models are considered (Chang & Ying, 1996) . As well as Continuous Entropy method is often the objective function for cognitive diagnostic CAT (Xu, Chang, & Douglas, 2005; Cheng, 2009). An e_ective evaluation should present items neither too difficult nor too easy for examinees (Lord, 1970), i.e., individuals would have approximately 50% probability to give the correct response for the item in the cases of dichotomous models. In the theoretical foundation of IRT model-based CAT, this fact is used as a synonym of matching difficulty levels of test items with an examinee's ability. However, 50% probability of correct response implies the equality of difficulty and ability parameters only in symmetric IRT models.

Nevertheless, in some cases, symmetric models are not appropriate for the item characteristic curve.

As practical examples, we can cite the 3 Parameter Logistic Model (3PL) which is implemented in Test of English as a Foreign Language (TOEFL; Educational Testing Service, 2015) and in the Armed Services Vocational Aptitude Battery (ASVAB; U.S. Department of Defense, 1984). In a more theoretical approach, Samejima (2000) proposed the asymmetric Logistic Positive Exponent (LPE) model, that generalizes the Logistic IRT models in order to solve the inconsistent relationship between the difficulties of items and the order of maximum likelihood estimates of ability.

CAT based on asymmetric IRT models can be criticized because maximizing Fisher and Kullback-Leibler information functions may result in administering items with too low or too high probability of success. Some simulation studies were done in this work in order to analyse selection methods based on Fisher (F), Kullback-Leibler (KL) Informations and Continuous Entropy (CE) under 3PL and LPE models and to verify what are the probability of correct response of the selected items (if they are close to 50%, according to Lord, 1970, or not).

In the simulation studies the a item parameter assumed values 0.3, 1 or 1.7, _ assumed values -2, 0 or 2, c parameter was equal to 0, 0.2 or 0.5, and, finally, if assumed values 0.5, 1 or 2. The results show all the three item selection criteria favor to administer items that are not at 50% correct response for the adopted asymmetric models, in contradiction to Lord (1970) idea that too difficult or too easy items should not be presented to an individual. For the 3PL model, the probability of correct response

of the selected item is closer to 50% adopting Fisher Information criteria than the other two methods. As well as the LPE model, the probability of correct response of the selected item is closer to 50% adopting CE method than the others.

Evaluating Effectiveness of Standard Error of Score Estimation as a Termination Criterion in CAT

Chansoon Lee, University of Wisconsin-Madison, Kyung (Chris) T. Han, GMAC

With its distinctive advantages, for example, improved measurement efficiency by administering test forms tailored to be most relevant to each individual, CAT is rapidly expanding its place in the education measurement field and is now widely used in various large scale assessments, small classroom exams, and even with self-paced learning courses. Like any other types of test, for test scores from CAT to be comparable across different test forms, among several other factors, the test forms should be equivalent in terms of score reliability, and that is usually ensured by controlling test information function (TIF) or standard error of θ estimation (SEE), which is simply an inverse of square root of TIF. In theory, if test forms result in the same SEE at the same θ point, they are expected to have the same measurement precision for people who are exactly at that θ point even if the test forms differ in test length, and that has been the basis of test developers who chose to allow test length to differ across test takers by applying CAT termination rules based on SEE. Several studies evaluated and compared different test termination rules for CAT, but there have been few studies that really focused on the relationship and interaction among controlled SEE, true measurement error, and test length (especially when test length is extremely short, for example, less than 10 items).

The main question of this study is about whether terminating CAT based on the SEE criterion really can ensure the measurement precision consistent under different θ estimation conditions for CAT with different test length. To answer this question, two different series of simulation studies are conducted. In study 1, several fixed test forms with different test length that all result in the same SEE at a θ point are simulated under extremely simplified and systematically controlled conditions with various θ estimation methods. In study 2, various CAT conditions with the SEE termination rule are simulated to systematically investigate the relationship among estimated SEE, true measurement error, and test length with different ability estimation methods for CAT. The findings of the two simulation studies suggest that SEE tends to be an over approximation of empirically observed conditional standard error of measurement when the test length exceeds about 10 items – in other words, using the SEE termination rule is usually safe to obtain or exceeds targeted measurement precision. When test length is shorter, however, relying solely on SEE often may not be enough to ensure targeted measurement precision depending on the choice of estimation method. The full paper includes comprehensive discussion on a link between controlling SEE and achieving the equivalent quality of test across different conditions and offers test developers and practitioners practical guidelines of successful administration of CAT programs with variable-test length.

PLATINUM SPONSORS

CAMBRIDGE ASSESSMENT

GMAC®

GRADUATE MANAGEMENT
ADMISSION COUNCIL

GOLD SPONSORS

PEARSON

ETS®

Listening. Learning. Leading.®

ASSESSMENT SYSTEMS
For Good Measure™

NCSBN
National Council of State Boards of Nursing

CiTO
now you know

UNIVERSITY OF
CAMBRIDGE
The Psychometrics Centre