IACAT 2019

10 - 13 June

International Association for Computerized Adaptive Testing

IACAT

Advancing the Science and Practice of Human Assessment

Conference Program and Abstracts

Graduate Hotel, University of Minnesota

Host: Assessment Systems Corporation

Welcome

Thank you for attending the 2019 conference of the International Association for Computerized Adaptive Testing (IACAT). This Program is primarily intended to provide the detailed conference schedule and presentation abstracts so that you can make the most of your visit. However, here are a few other details:

Location

The conference is being held at the Graduate Hotel on the University of Minnesota campus. Guest of the Graduate hotel receive complimentary bike rentals and free passes to the University's Recreation and Wellness Center. Check in with the front desk for more details.

Transportation will be provided for excursions.

Travel

For more information on how to get to the Twin Cities of Minneapolis & St. Paul, and then to the hotel, visit http://iacat.org/2019-venue-and-travel.

Welcome messages from IACAT President and Vice President

Learn about the importance of this conference from Dr. Mark Reckase and Dr. Alina von Davier at http://iacat.org/presidents-welcome.

Registration

You must register online for the conference, please do so at http://iacat.org/register-2019-conference.

Visiting the Twin Cities

If you want to see some of the Twin Cities before or after the conference, some suggestions are at http://iacat.org/about-twin-cities .

Sponsors

We are grateful for our sponsors! If you see someone that works with the organizations below, please thank them.

Organizing Committee:

Nathan Thompson, Assessment Systems Corporation (Host)
David J. Weiss, University of Minnesota
Alina von Davier, ACTNEXT by ACT
Mark Reckase, Michigan State University
Alper Sahin, METU Northern Cyprus Campus
John Barnard, EPEC

Monday, June 10

Hours	Name(s)	Title	Room
8:30 - 10:00	John Barnard & David J. Weiss	Workshop: Introduction to Item Response Theory and Computerized Adaptive Testing	Summit
8:30 - 10:00	Theo Eggen & Angela Verschoor	Workshop: Simulations and CAT	Pathways
8:30 - 10:00	Chun Wang	Workshop: Multidimensional CAT	Think 5
Break	---------- COFFEE ----------		Think Tan k
10:15 - 11:45	John Barnard & David J. Weiss	Workshop: Introduction to Item Response Theory and Computerized Adaptive Testing	Summit
10:15 - 11:45	Theo Eggen & Angela Verschoor	Workshop: Simulations and CAT	Pathways
10:15 - 11:45	Chun Wang	Workshop: Multidimensional CAT	Think 5
11:45 - 12:30	Lunch	(on your own)	
12:30 - 12:40	Nathan Thompson & David J. Weiss	Host Welcome	Meridian 234
12:40 - 12:50	Mark Reckase	Presidential Welcome	Meridian 234
12:50 - 1:50	Alina von Davier	Incoming Presidential Keynote: The Road Ahead: From Computer Adaptive Testing to Artificially Intelligent Measurement	Meridian 234
Break	(JUST TO CHANGE ROOMS)		
2:00 - 3:00	Ricardo Primi, Mario Piacentini, & Tomoya Okubo	Exploring the use of deep learning to score tasks of PISA 2021 assessment of creative thinking	Summit
	Alan Mead & Sheng Zhang	Towards a strong AIG CAT: A review of three feasibility studies	Summit
	Catherine Close	Exploring Adaptive Measurement of Change for Individual Student Progress Monitoring	Pathways

	Jonghwan Lee, Christy Schneider, Sylvia Scheuring, Sukkeun Im, & Jungnam Kim	Running Simulations to Maintain Score Compatibility of CAT Across Years	Pathways
	David Shin & Yuehmei Chien	A Visual Multistage Module and Panel Assembly Tool	Think 5
	Angela Verschoor	An ATA model for multistage testing	Think 5
	Juntao Wang & Gongjun Xu	Sequential Gibbs Sampling Algorithm for Attributes Profiles in DINA	Meridian 1
	Kang Xue	A Deep Feedforward Network based Semi-Supervised Learning Method to Improve the Performance of Diagnostic Classification	Meridian 1
Break	---------- SNACKS ----------		Meridian Foyer
3:20 - 5:00	Michael Yudelson, Ada Woo, Stephen T. Polyak, Lu Ou, Yigal Rosen, & Ilia Rushkin	**Symposium: Learning meets Assessment: Adaptation and Personalization at Scale for Practitioners** Integrating Learning, Measurement, and Navigation: Introduction of a Recommendation and Diagnostic (RAD) API *Ada Woo ACTNext by ACT, Inc.)* *Stephen T. Polyak ACTNext by ACT, Inc.)* *Lu Ou (ACTNext by ACT, Inc.)* The Effects of Adaptive Learning in a Massive Open Online Course on Learners' Skill Development *Yigal Rosen (ACTNext by ACT, Inc.)* *Ilia Rushkin (Harvard University)* Building and Picking the Right Model for Learning and Assessment. Notes for Model Developers *Michael V. Yudelson (ACTNext by ACT, Inc.)*	Meridian 1
	Angela J. Verschoor, Stéphanie Berger, Patrick Meyer, Theo J. H. M. Eggen, Urs Moser, Jakob Wandall, & Gage Kingsbury	**Symposium: Development and Validation of three Vertical Scales for Formative Computer Adaptive Testing** Mindsteps (Switzerland) *Stéphanie Berger, Angela J. Verschoor, Theo J. H. M. Eggen, & Urs Moser* Nationale test og folkeskolens digitale prøver (Denmark)	Pathways

		Jakob Wandall The RIT scales by NWEA (USA) *Gage Kingsbury, Patrick Meyer*	
	Hua-Hua Chang, Nathan Thompson, Shengyu Jiang, Chun Wang, Shiyu Wang, & Yinhan Chen	**Symposium: From Adaptive Testing to Artificial Intelligence and Smart Learning** AI and Machine Learning in Psychometrics? Old news. *Nathan Thompson* On-The-Fly Parameter Estimation Based on Item Response Theory in Adaptive Learning Systems *Shengyu Jiang & Chun Wang* Automated Attribute Hierarchy Detection with Application to Adaptive Learning *Shiyu Wang & Yinhan Chen* Understanding Interactive Items' Characteristics by Deep Learning-based Process Data Analysis *Susu Zhang, Xueying Tang, Zhi Wang, Jingchen Liu, & Zhiliang Ying*	Summit
5:00 - 6:00	Welcome Reception	Hors D'oeuvers and Networking	Pinnacle Ballroom
6:00 PM	River Cruise	Loading bus 600-615; bus ride 615-645; load boat 645-700; cruise 7:00-10:00	

Tuesday, June 11

Hours	Name(s)	Title	Room
7:30 - 8:00	Breakfast		Meridian 234
8:00 - 9:30	Jonghwan Lee, Sylvia Scheuring, Sukkeun Im, Jungnam Kim, & Christy Schneider	Relaxing Constraints for Better Measurement Precision during a CAT	Summit

	(open due to cancellation)		Summit
	Angela Verschoor	On the impact of measurement error in computerized adaptive testing	Summit
	Bingnan Jiang	Enhancing Effectiveness of CAT Simulation with the Shadow-Test Approach in RSCAT	Pinnacle
	Chen Li & Michael Chajewski	Penalization to Item Selection Inadequacy in a CAT	Pinnacle
	Catherine Mintz, Deborah Harris Ph.D., & Aaron McVay	An Investigation of Item Selection Criteria for Passage-based Items	Pinnacle
	James B. Olsen	Validity Evidence from a Formative CAT Educational Mathematics Assessment	Think 5
	William Insko Jr.	Interval Validation Method: Achievement Level Setting Based on Large Item Pools Used in Computerized Adaptive Testing	Think 5
	Shuqin Tao, Luciana Cançado, & Brett Morrow	Selecting off-grade items appropriately in a vertically-scaled adaptive test	Think 5
	Andreas Frey, Aron Fink, & Christian Spoden	Consideration of Item Position Effects in CAT with the Continuous Calibration Strategy	Meridian 1
	J. B. Weir	Enemy Item Detection Using Data Mining Methods	Meridian 1
	Haruhiko Mitsunaga	CAT system utilizing automated procedure for detecting enemy items based on NLP	Meridian 1
9:30 - 9:50	Break	---------- COFFEE/SNACKS ----------	Meridian Foyer
9:50 - 10: 50	Jimmy de la Torre	Keynote: Recent advances in cognitive diagnosis computerized adaptive testing	Meridian 234
10:50 - 11:00	Break	(JUST TO CHANGE ROOMS)	
11:00 - 12:00	Audra Kosh	Trends in Rapid Guessing Behavior Within a Low-Stakes K-12 Computer Adaptive Test	Summit
	Yanyan Fu & Chris Han	A Comparison of Three Approaches That Permit Answer Review and Revision in Computerized Adaptive Testing	Summit

	John Hansen & Jerel Unruh	Measuring Nonverbal IQ in Autism: Challenges and Opportunities Building a Measure and Internet Administration Application	Pinnacle
	Mo Chen, Kenneth Poon, Yong Hwee Nah, Huichao Xie, Vahid Aryadoust, & Nicolette Waschl	Developing and Piloting an Item Bank for a Computerized Adaptive Test of Adaptive Behavior Skills	Pinnacle
	John Barnard, Gage Kingsbury, Nathan Thompson, Anthony Zara	A History of CAT	Think 5
	(open due to cancellation)		Think 5
	Seyedahmad Rahimi, Valerie Shute, Russell Almond	The Technical Underpinning of Physics Playground	Meridian 1
	Benjamin Deonovic	An Urn-scheme to Track Accuracy and Response Times in Learning Environments	Meridian 1
12:00 - 12:50	Lunch		Meridian 234
12:50 - 2:30	Alper Sahin, Eren Can Aybek, Murat Doğan Şahin, Selahattin Gelbal, Ebru Balta, Arzu Uçar	**Symposium: CAT in Turkey: From Past to Present** CAT in Turkey: An Introduction *Alper Şahin* Development of the CAT version of a Turkish Interest Inventory *Eren Can Aybek, R. Nükhet Çikrikçi* Developing a MCAT to Assess Students' Language Competencies *Murat Doğan Şahin & Selahattin Gelbal* Detection of cheating behavior in online unproctored CATs via a validation test *Ebru Balta, Arzu Uçar, Alper Şahin* The Development of Theoretical Knowledge Test for Driving License Exam as a Computerized Adaptive Test *Nükhet Çıkrıkçı, Seher Yalçın, İlker Kalender, Emrah Gül, Cansu Ayan, Gizem Uyumaz, Merve Şahin Kürşad, Ömer Kamış*	Meridian 1

	Michelle Barrett, Angie McAllister, Benjamin Deonovic, Krista Mattern, & Drew Hampton	**Symposium: Adaptive Learning and Adaptive Assessment: Intersections and Synergy** Design framework: Definitions of "adaptive" in learning and assessment systems and algorithmic measures of success *Angie McAllister & Michelle Barrett* Theoretical connections: Statistical models for adaptive learning and adaptive assessment: *Benjamin Deonovic* Making it real: Open source tools (e.g., GIFT, OLI, RSCAT, catR, mstR) and industry standards for adaptive learning and assessment systems ((IEEE) AIS, LTSC, xAPI, (IMS) Caliper, QTI CAT) *Drew Hampton & Michelle Barrett* Knowing it works: Reflections on measuring efficacy of adaptive assessment and learning systems at scale *Krista Mattern*	Pinnacle
	Laurie Davis, Can Shao, Bess Patton, & Logan Rome	**Symposium: Identifying Rushing in the i-Ready Diagnostic Computerized Adaptive Test** Developing Item-Level Rush Flags for the i-Ready Diagnostic Assessment *Bess Patton, Logan Rome* Developing Test-Level Rush Flags for the i-Ready Diagnostic Assessment *Logan Rome, Bess Patton* Comparison of Three Methods for Detecting Test-Level Rushing in the i-Ready Diagnostic Assessment *Can Shao, Logan Rome*	Summit
Break	(JUST TO CHANGE ROOMS)		
2:40 - 3:40	David J. Weiss	Keynote: The CAT Curmudgeon: Some Thoughts from 50 Years of CAT	Meridian 234
Break	---------- SNACKS ----------		Meridian Foyer

4:00 - 5:00	Jordan Stoeger, Tianjiao (Tina) Yin, & Xiaoqiu Xu	Getting Started with CAT: A Case Study	Summit
	Ryan A Wilke & Cody Diefenthaler	Is Your Item Pool Ready for the Demands of CAT?	Summit
	Mariana Curi, Thales Akira Matsumoto Ricarte, & Alina Von Davier	The MST Role to Reduce Public Cost in Testing	Pinnacle
	Xiao Luo	A Top-down Design Paradigm of Computerized Adaptive Multistage Test	Pinnacle
	Kuo-Feng Chang & Won-Chan Lee	Dual-Objective Adaptive Testing Using High-Order Item Response Theory Models	Think 5
	Yikai Lu, David J. Weiss, & Chun Wang	Multidimensional CAT Measuring Patient-Reported Outcomes in a Hospitalized Population	Think 5
	Stanley Rabinowitz	Implementing a Large-Scale Computer Adaptive Assessment System: The Role of Culture and History	Meridian 1
	Tetsuo Kimura	Test-takers' Psychological Aspects in Case of a Small-scale Multistage Computer Adaptive Testing	Meridian 1
5:00 - 7:30	Sculpture Garden	Loading bus 500-515; bus ride 515-530; art 530-700; bus ride home 700-715	
	Dinner	On your own	

Wednesday, June 12

Hours	Name(s)	Title	Room
7:30 - 8:30	Breakfast		Meridian 234
8:30 - 9:30	Po-Hsi Chen & Chia-Ling Hsu	The Development and Efficiency of the Computerized Adaptive Testing Version for the Test of Chinese as a Foreign Language	Summit
	Dee Kanejiya	Conversational Assessments for AI based Adaptive Testing	Summit
	Alan Mead & Jilin Huang	A machine learning "Rosetta Stone" for psychologists and psychometricians	Pinnacle

	Nathan Thompson & Alper Sahin	Feasibility of Using Small Samples to Train an Automated Essay Scoring Engine	Pinnacle
	Eva K. Fenwick, John Barnard, Aiden Loe, Alfred Gan, Jyoti Khadka, Konrad Pesudovs, Shu Yen Lee, Gavin Tan, Tien Y. Wong, & Ecosse L. Lamoureux	Psychometric evaluation of computerized adaptive tests to assess quality of life in people with diabetic retinopathy	Think 5
	Ryan EK Man, Bao Sheng Loe, Jyoti Khadka, Gwyn Rees, Eva K Fenwick, & Ecosse L Lamoureux	Evaluation of a Computerized Adaptive Testing System for the Impact of Vision Impairment Questionnaire (IVI-CAT)	Think 5
	Xue Zhang & Chun Wang	Assessing item-level fit within the multistage testing environment	Meridian 1
	Maaike van Groen	Using Multidimensional Item Response Theory for the Multistage End of Primary School Test	Meridian 1
	Break	---------- COFFEE/SNACKS ----------	Meridian Foyer
9:50 - 10: 50	Chun Wang	Keynote: Multidimensional Computerized Adaptive Testing in Health Measurement: Lessons Learned	Meridian 234
	Break	(JUST TO CHANGE ROOMS)	
11:00 - 12:00	Duke Sowunmi & H.O.Owolabi	Adaptive essay test assessor for secondary school economics	Summit
	(open due to cancellation)		Summit
	Chia-yi Chiu & Yuan-Pei Chang	Advance in CD-CAT: The General Nonparametric Item Selection Method	Pinnacle
	Soleyman Zolfagharnasab	A Nonparametric approach to CAT	Pinnacle
	Ye (Cheryl) Ma, Johnny Denbleyker, & Shuqin Tao	Scripted CATs via an Adaptive Blueprints with Multistage Considerations: A simulation study	Think 5
	(open due to cancellation)		Think 5
	Jinah Choi & Windy Torgerud	Evaluating model fit and skill interdependency of granular growth model in a formative assessment system	Meridian 1
	Yeow M. Thum	Realigning the Scale of One Item Pool to Another: IRT Linking Using Growth Data	Meridian 1

12:00 - 12:50	Lunch		Meridian 234
12:50 - 2:30	Mark Reckase, Andres Paez, Sewon Kim, & Unhee Ju	**Symposium: Converting from Paper-and-Pencil to CAT: Developing CAT Designs to Meet Required Constraints** PreSaber 11 Design and Uses *Andrés Páez* Designs of CAT Versions of the PreSaber 11 Components *Sewon Kim* Item Pool Designs and CAT Design Evaluations *Unhee Ju* Final CAT Designs and their Properties *Mark D. Reckase*	Pinnacle
	Alina von Davier, Jingchen Liu, Viktor Qvarfordt, Geoff LaFlair, & Meirav Arieli-Attali	**Symposium: Digital Adaptive Learning and Assessment Systems** An Exploration of Process Data in Computer-based Assessment *Jingchen Liu* Sana Learn: A Personalized Learning Recommender System *Viktor Qvarfordt* A Machine-Learned Construct of Language Proficiency and an Adaptive Language Assessment *Geoffrey T. LaFlair* Developing a Learning and Assessment System with the expanded ECD Framework; the HERA Showcase *Meirav Arieli Attali*	Summit
	David J. Weiss, Jieun Lee, Chaitali Phadke, Chun Wang, King Yiu Suen, & Matthew Finkelman	**Symposium: Adaptive Measurement of Change (AMC): Identifying Psychometrically Significant Change One Examinee at a Time** Change in K-12 reading achievement on two occasions *Jieun Lee & David J. Weiss* Application of Omnibus Hypothesis Tests to K-12 Math Data in Measuring Growth *Chaitali Phadke & David J. Weiss*	Meridian 1

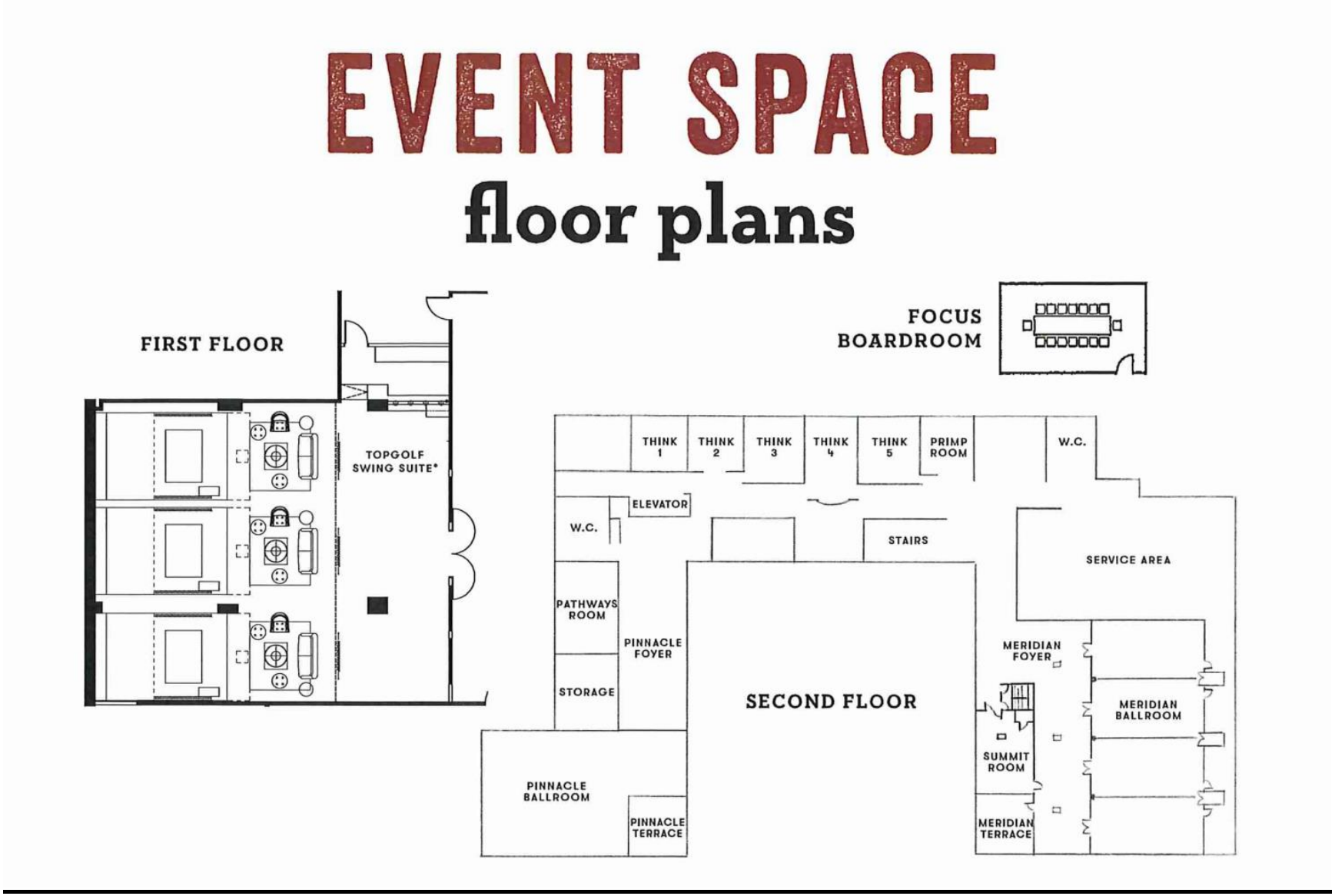
		Change in multiple patient-reported outcomes across multiple occasions *King Yiu Suen, Chun Wang, & David J. Weiss* Time efficient adaptive measurement of change *Matthew Finkelman & Chun Wang*	
	Break	(JUST TO CHANGE ROOMS)	
2:40 - 3:40	Shiyu Wang (Early Career Award)	Keynote: Computerized Adaptive Testing with Response Revision: Challenges, Solutions and Applications	Meridian 234
	Break	---------- COFFEE/SNACKS ----------	Pinnacle Foyer
4:00 - 5:00	Darrin Grelle	An Evaluation of Collapsing Response Categories for Innovative Question Types Using the GPCM	Summit
	Xin Lucy Liu, Xuechun Zhou, & Haiqin Chen	Impact of Innovative Item Scoring on Constrained Computerized Adaptive Testing	Summit
	Mingjia Ma & Terry Ackerman	The Influence of Dimensionality on Item Exposure Rate under CAT Administration	Pinnacle
	(open due to cancellation)		Pinnacle
	Xiaowen Liu & H. Jane Rogers	Detecting Aberrant Behavior in Computerized Adaptive Testing: The Lognormal Response Time Model	Think 5
	Luz Bay	Effects of Compromised Items in Computer-Adaptive Tests: A Retrospective Study	Think 5
	King Yiu Suen & David J. Weiss	Effect of Routing Errors on the Psychometric Properties of Multistage Tests	Meridian 1
	Jing Lu, Chun Wang, & David Weiss	Using response time to improve precision and efficiency of computerized adaptive testing	Meridian 1
5:30 - 8:00	Conference Dinner		Meridian 234

Thursday, June 13 (Sessions to be held at Recreation Center next to Hotel)

Hours	Name(s)	Title	Room
7:30 - 8:30	Breakfast		Pinnacle
8:30 - 10:00	Giada Spaccapanico Proietti, Mariagiulia Matteucci, Stefania Mignani, & Bernard Veldkamp	Chance-constrained test assembly	Beacon
	Tyler Matta	A tree-based algorithm for test blueprint creation	Beacon
	Emre Gonulates	A Search for a Practical Definition of an Optimum Item Pool	Beacon
	Zhongmin Cui	On Measuring Adaptivity of an Adaptive Test	MP 1
	Adam E. Wyse & James McBride	A Framework for Measuring the Amount of Adaptation of Rasch-based CATs	MP 1
	G. Gage Kingsbury & Steve Wise	Three Measures of Test Adaptation Based on Optimal Test Information	MP 1
	Fernando Austria Corrales, Dafne Saavedra, & Nathan Thompson	10/66 Dementia Research Group Cognitive Test Battery: results from IRT and CAT-simulation studies	MP 2
	Kathryn Jackson, Ben Schalet, Michael Kallen, Aaron Kaat, Michael Bass, Richard Gershon, & David Cella	Application of CAT Stopping Rules to Increase Precision and Decrease Burden for Health Outcomes Research	MP 2
	Aaron Kaat & Richard C. Gershon	Examining Exposure Control on the NIH Toolbox Picture Vocabulary Test	MP 2
	Beyza Aksu Dunya	Investigating Item Parameter Drift As a Problem of Ability Parameter Drift in CAT	MP 6
	Nixi Wang & Chun Wang	Exploring Differential Item Functioning in Cognitive Diagnostic Computer Adaptive Testing	MP 6
	(open due to cancellation)		MP 6
	Break	---------- COFFEE/SNACKS ----------	Beacon
10:20 - 11: 20	Dong Gi Seo & Sun Huh	Application of Computerized Adaptive Testing using Medical Class Examination in South Korea	Beacon

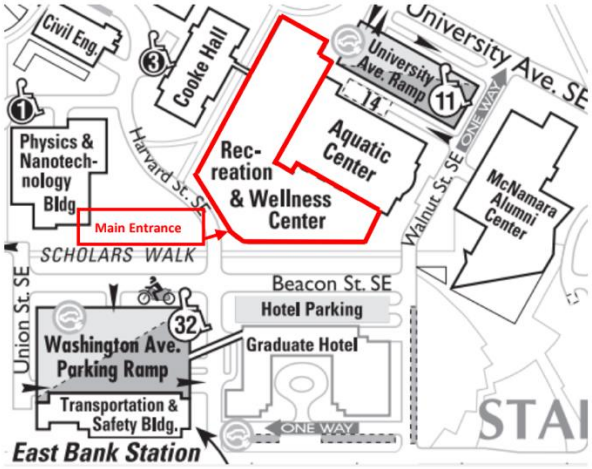
	(open due to cancellation)		Beacon
	Stephanie Varga	MYMAP: A Graph Theory Approach to Modelling Attribute Mastery	MP 1
	John Denbleyker, Shuqin Tao, Ye Ma & Mingqin Zhang	Evaluation of Content Maps Constructed via a K12 Computer Adaptive Mathematics Assessment	MP 1
	Emanuela Botta	Experimentation of a MST for math skills in large scale surveys in Italy	MP 2
	Dimiter M. Dimitrov, Hanan M. ALGhamdi, & Abdullah A. Alqataee	Multistage Testing Using the D-Scoring Method: Research and Piloting at the NCA in Saudi Arabia	MP 2
	Scott B Morris, Michael Bass, Matthew Lauritsen, & Richard Neapolitan	A Predicted Error Reduction Stopping Rule for Multidimensional Computer Adaptive Tests	MP 6
	Zhuoran Wang, Chun Wang, & David J. Weiss	Adaptive Multiclassification Testing with Multidimensional Polytomous Items	MP 6
	Break	(JUST TO CHANGE ROOMS)	
11:30 - 12:30	Steve Wise	Keynote: Expanding the Meaning of Adaptive Testing to Enhance Validity	Beacon
12:30 - 12:45	Alina von Davier, Nathan Thompson, & David J. Weiss	Closing Remarks	Beacon

The Graduate Hotel:

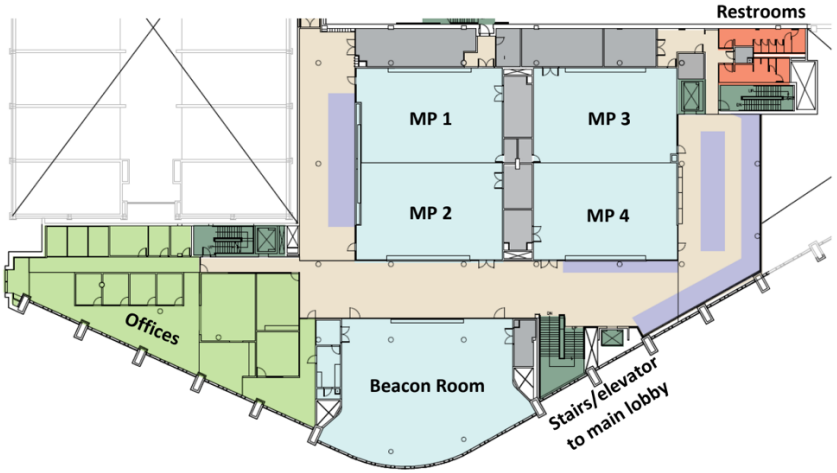


The Recreation and Wellness Center:

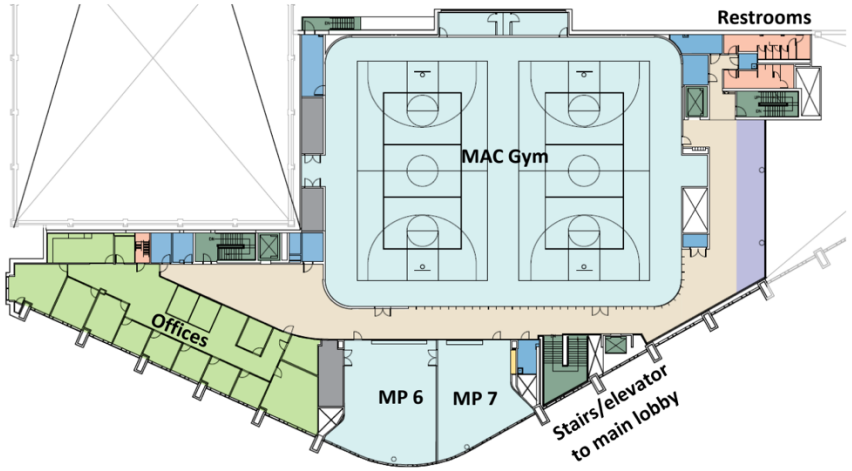
Recreation and Wellness Center Main Entrance



Recreation and Wellness Center 2nd Floor



Recreation and Wellness Center 4th Floor



Abstracts

Exploring the use of deep learning to score tasks of PISA 2021 assessment of creative thinking

Authors
Ricardo Primi (Universidade São Francisco Edulab21 Instituto Ayrton Senna, Campinas, Brazil)
Mario Piacentini (OECD, Education and Skills Directorate, Paris, France)
Tomoya Okubo (Research Division, The National Center for University Entrance Examinations, Tokyo, Japan)

Abstract
Creativity has been recognized as one of the most important 21st century skills. However, the capacity to think creatively is rarely included in assessments that monitor the performance of educational systems. Research in psychology has traditionally measured creative thinking potential via divergent thinking tasks. These tasks ask subjects to produce original and valuable ideas in response to an open prompt. Responses are then scored by raters and aggregated to produce a creativity score. But this method faces a number of challenges. One of the most important ones is the burden it imposes on raters. The most common procedure to score divergent thinking tasks with a sufficient degree of reliability has been in fact to ask multiple raters to score each response. The high cost of scoring these tasks has limited their inclusion in large-scale assessments. Reliable, automated coding systems can reduce these costs and thus encourage the development and use of tests of creative potential in large-scale data collections. The Programme for International Student Assessment (PISA) will face the challenge of identifying a reliable and feasible method for scoring the assessment of Creative Thinking, the PISA innovative domain for 2021. In this paper, we present the results of an exploratory study on the use of artificial neural networks to score prototype tasks that have been developed for the new PISA assessment. In the data included in this paper, 202 15-year-old students from South Africa responded to four tasks out of a pool of 26. The tasks are meant to assess three facets of creative thinking as defined in PISA (the capacity to generate ideas that are sufficiently diverse, the generate to elaborate an original idea, and the capacity to evaluate and improve on ideas) within three domains of application (written expression, social problem solving, scientific problem solving). Out of a sample of 9944 responses, 8154 responses were used for training the algorithm and 1790 responses for validation. We trained a two-layer neural network with word embeddings. In a preliminary iteration of the automated scoring procedure, we achieved a correlation of 0.48 between the average score provided by the two human raters and the score predicted by the model. The paper discusses possible ways to further improve this system in order to generate deep learning models that can effectively score creative thinking tasks.

Towards a strong AIG CAT: A review of three feasibility studies

Authors
Alan D. Mead (Talent Algorithms Inc.)
Sheng Zhang (Illinois Institute of Technology)

Abstract
Automated item generation (AIG; Irvine & Kyllonen, 2002; Gierl & Haladyna, 2013) and computerized adaptive testing (CAT; van der Linden & Glas, 2010) were innovations arising from the application of computers to improve psychological measurement (Haladyna, 2013). By some measures, the ideal innovation would be an "AIG CAT" where the CAT generates an item with the desired difficulty on-the-fly during testing.
This abstract summarizes two feasibility studies and presents a new, third feasibility study. These studies are part of a research program to produce a CAT assessment that generates items automatically during the assessment using strong AIG.
Weak AIG (Drasgow, Leucht & Bennett, 2006) includes template-based approaches and has two distinguishing features: (1) items generated from a single template tend to be similar in content and difficulty; and (2) little is known about why an item is easy or difficult.
Strong AIG is better suited to an AIG CAT in which the CAT generates the desired item with specific content and difficulty parameters. The assessment discussed in this abstract uses verbal reasoning items, for which there are some theoretical understanding (Whitely, 1976; Sternberg, 1977; Duran, Enright, & Pierce, 1987; Jones & Estes, 2015).
Study 1: Psychometics of AIG items
The first feasibility study investigated the effectiveness and comparability of AIG for verbal reasoning. Three 20-item AIG forms were compared to an SME-written form. The AIG forms showed better coefficient alpha reliability estimates (0.72 and 0.73 vs. 0.57) and comparable convergent and criterion-related validity coefficients. The AIG forms measured the same construct as the SME-written form with better reliability and comparable validity.
Study 2: Examinee reactions to flawed items
One of the dangers of a CAT AIG is the near certainty that some examinees will encounter a poorly-generated item. In the second study, flawed items were intentionally generated and inserted onto exam forms. Examinee reactions to the exams and performance were compared across fourbetween-subjects conditions: control (no intentionally flawed items), 10% flawed items, 20% flawed items, and 40% flawed items. About 25 volunteers were recruited for each group from Amazon MTurk.
We found that even when 40% of the exam was composed of items intended to be flawed, this had very slight, mostly non-significant effects on examinee reactions and performance. The mean Cohen's d was -0.11 and only the difficulty reaction facets showed a significant effect (and only the 40% v. control had a significant post hoc test).
Study 3: Predicting AIG item difficulty

The final study, which will be the majority of the presentation, will describe preliminary analyses of a model of item difficulty that can be used to target an AIG CAT.

Exploring Adaptive Measurement of Change for Individual Student Progress Monitoring

Author
Catherine Close

Abstract
Student learning - expressed as growth – is a primary focus of instruction. Commonly, growth is summarized at the group level: classroom, school, district and so forth. While such summaries are necessary and continue to inform group instruction, it is also known that students progress at different rates. Any such individual differences are often masked by aggregate summaries of growth. Hence, to better personalize instruction and meet individual students' needs, we must reliably measure how each student is progressing with instruction. Typically, growth is assessed in relation to meeting a target benchmark such as, no longer needing intervention or being on track to be proficient on grade-level expectations by the end of the school year.
Various test types and methods of quantifying individual student growth have been in use for a long time. Until recently, however, there was over-reliance on static tests such as paper and pencil tests, tests that might not be truly informative for all students. Once students had tested multiple times, growth models such as the difference scores approach (difference between time 1 and time 2 scores), were then used to quantify growth over time.
Extant literature suggests that these types of tests and growth models are fraught with issues that complicate individual student progress monitoring. The use of the adaptive measurement of change approach (AMC) that relies on computerized adaptive tests (CAT) has been suggested for reliably quantifying individual student growth over time to provide actionable progress monitoring data (Kim-Kang and Weiss, 2008; Finkelman et al., 2010).
This study examined the use of a computerized adaptive test (CAT) interim assessment- Renaissance Star Reading- for measuring growth. Because Star Reading is tailored to each student, it delivers a unique test that is matched with the student's ability level. This study uses real data from a constrained CAT which has not been studied before. The use of such an adaptive test to measure growth is contrasted with the more traditional fixed-form tests. In addition, the adaptive measurement of change (AMC) approach is presented as a practical and reliable method that can be used to show individual student growth against a backdrop of other individual growth measurement models. The AMC approach takes the inherent uncertainty in all scores into consideration hence separating growth from noise caused by measurement error to show true growth. Preliminary findings indicated that the AMC method provides a practical, reliable, and valid approach to measuring each student's progress throughout the school year even when tests are constrained to balance content and limit item exposure.

Finally, while the choice of the test type and the growth model is important in the individual measurement of change, some important considerations such as, the nature of the item pool, test constraints, and person fit are also discussed in terms of how they may impact student growth.

Running Simulation to Maintain Score Compatibility of Adaptive Tests Across Years

Authors
Jonghwan Lee
Christy Schneider
Sylvia Scheuring
Sukkeun Im
Jungnam Kim

Abstract
Conducting a simulation study on an educational statewide assessment prior to the operational computer adaptive test (CAT) serves two main purposes from a technical perspective: (a) the study results allow the state to determine if the item pool is sufficient to find a feasible set of items for students within grade who have different ability levels, and (b) the study results allow the state to evaluate the functioning of the engine's item selection algorithm to ensure that the state's construct for test scores (e.g., ELA proficiency) is being represented as intended. Test scores represent how students perform regarding the test construct, and administering blueprint-compliant items ensures that the test scores have appropriate representativeness of each strand (or construct).
A fundamental consideration when measuring students across time is that the item pool attributes that underlie student test events remain the same or stay parallel. When item pools are parallel across time, predicted **student score distributions** are expected to stay the same even as items are refreshed or removed (Stocking, 1994). Davey, Pitoniak, and Slater (2016) noted that item pools should be assembled like an expanded version of the test form with comparable criteria to fixed forms. Barnard (2015) wrote that many CAT item banks are considered sufficiently parallel when the number of items in the blueprint cells are comparable and information functions for the pools overlay. These guidelines suggest that score comparability can be monitored through simulations by investigating predicted score distribution. This is critical when item pools need to be adjusted over time to better target student ability, which is not uncommon in educational assessments. This simulation study uses bias, RMSE, item exposure rate, SEM, and reliability analyses to investigate the score distributions and classification across two item pools of different difficulty and the same constraints for Grades 3–8 ELA and Mathematics. Findings show highly similar results and support for using the more difficult item pool.

A Visual Multistage Module and Panel Assembly Tool

Authors
David Shin, Ph.D. (Pearson, Principal Research Scientist)
Yuehmei Chien

Abstract
In cases where MST may become a practical and possibly beneficial alternative to linear fixed forms, the decision must be carefully considered from different aspects because the transition is neither trivial nor inexpensive. Making such decision needs research that requires resources and time. Is there an efficient way to facilitate decision making that can be adapted to different existing linear tests for different testing programs? Also, If MST is a good choice, how to quickly settle on one MST design that is optimal based on evidence? Lastly, how can the research results be easily understood by the stakeholders and decision-makers who might not have solid psychometric background?
To answer the questions above, we build necessary tools, called the MST suite, that provide all we need to solve those issues at once. The MST suite must allow the test design to be easily specified, as well as the content and statistics constraints. It must efficiently assemble modules/panels for the given test design, diagnostically analyze the pool, and generate visual and interactive reports. There are four main components in the MST suite briefly described below:
1. The Module Assembler
The Module Assembler includes a friendly interface that allows the user to easily specify the test design including the MST structure (i.e. the number of panels, stages, modules), along with various content and statistical constraints. The test design is then converted to a Linear Program (LP) format, which are linear inequality equations that a mixed integer solver can process into a complete and optimal solution. If the the LP problem is infeasible, diagnostic information may be provided.
2. The Item Bank Analyzer
Item bank quality is essential to MST and often directly impacts the decision. The Item Bank Analyzer provides analytic functions for deeper understanding of the compositions of the item bank and for helping to maximize measurement precision while using the bank efficiently. These analytic functions include content structure analysis, item quality analysis (e.g. item difficulty, discrimination, etc.), and other item attribute analysis (e.g. item usage and average response time.)
3. The MST Simulator
The MST Simulator has functionalities that are often found in other CAT simulators. The primary difference for MST is the routing rule, which adaptively routes the test taker to the next stage wherever there is a branch on the path. The routing rules provided include: cut scores in IRT ability scale, cut scores in IRT expected number correct score scale, and maximum Fisher information. Also, multiple panels are allowed for item exposure control and different simulation factors are able to specified for a comparison study.
4. The Visual Generator
The Visual Generator presents the MST modules and simulator results in different views through filterable tables and interactive plots. Through visual observation, decision makers can easily understand assembled modules/panels and effectively compare various MST designs or testing modes.

The MST suite will be demonstrated through a simulation study using a real operational item bank. The simulation results will be compared to fixed forms, previously built from the same pool.

An ATA Model For Multistage Testing

Author
Angela Verschoor (Cito)

Abstract
Despite an already long tradition in Multistage Testing (MST), the construction of one still remains an art: decisions regarding stages, composition of modules and routing that have to be taken are usually based on simple rules of thumb, gut feelings or previous experience. On the other hand, Automated Test Assembly (ATA) provides an excellent framework for many decisions to be optimized in a systematic way: which combination of items fulfills all specifications but still provides the most accurate measurement? Unfortunately, all ATA models devised until now only regard linear tests.
Even for relatively simple situations, questions like “At what length of the first stage in a two-stage test will the measurement error be optimal?” will yield varying answers from experts, while clearly only one answer could be correct.
In this paper, we present an ATA model for MST. The model user only needs to specify a limited set of specifications: next to the “standard” requirements for linear testing (content restrictions, practical considerations, etc.), the model assumes only an outline of the desired MST design: a number of stages, and a number of modules per stage). The other decisions (selection of items into the modules, routing rules) will be optimized in the model. For the objective function, two possibilities are offered: the first objective function assumes a flat threshold for the Fisher information function over a user-defined interval, while the second objection function minimizes the Root Mean Squared Error for a target population. As the model is non-linear, standard LP-approaches to solve these models might be cumbersome. Therefore, local search methods like Genetic Algorithms or Simulated Annealing seem to be more appropriate for this class of models. A very simple local search method will be presented, providing optimal or near-optimal results in short time.
Although results are heavily dependent on the exact constraints and available item pool, the model shows that in general in a two-stage test a relative short first stage will outperform a test with a longer first stage. Similarly, a 1-3-3 MST will in general outperform a 1-2-4 MST.

Sequential Gibbs Sampling Algorithm for Attributes Profiles in DINA

Author
Juntao Wang
Gongjun Xu

Abstract
We focus on the MCMC algorithm for the Deterministic Inputs, Noisy "and'' Gate (DINA) model. When the number of attributes K is large, MCMC for DINA maybe slow. For each attribute profile, 2^K calculations are needed to get full conditional distribution. We proposed sequential methods to sampling entries of attribute profiles one by one where it only needs to calculate $2 \times K$ times for each attribute profile. We proposed two different forms of the sequential methods: (1) the sequential I; it assumed attribute entries are generated from prior in a given sequence, (2) the sequential II; it assumed any attribute entry is generated conditional on population parameters and all other entries. Both of the sequential methods can help get closed forms of full conditional distributions, and sample attribute entries one by one. The new methods for DINA can enhance the computing speed of MCMC. For large K, we proposed a proper configuration of prior to get a sparse latent classes. To show the estimation performance of the new methods, existed methods such as the R package "CDM"(Rupp and van Rijn, 2018) and software "JAGS"(Zhan et al., 2019) are compared to our methods. In mild situations(K is smaller than 10), our methods will obtain comparable estimation. In extreme situation($K = 15$), "CDM" and "JAGS" doesn't converge or works slowly, however, our methods can still work.

A Deep Feedforward Network based Semi-Supervised Learning Method to Improve the Performance of Diagnostic Classification

Author
Kang Xue

Abstract
When analyzing a particular assessment dataset, inappropriate diagnostic classification models (DCMs) impacts the classification accuracy and parameter estimation. Although the existing research showed that, generally, the DINA, attribute hierarchy model (AHM), and rule space model (RSM) were mostly used with math; the reduced RUM, and general models (e.g. G-DINA, LCDM) were mostly used with reading, selecting an appropriate DCM for a particular assessment still takes lots of research effects.
In contrast to existing methods which tried to improve classification performance by adding new parameters or rebuild statistic models, in this paper the misclassification obtained by an inappropriate DCM is viewed as noisy or incomplete labels for examinees. According to this novel point of view, a Deep Feedforward Network (DFN) based semi-supervised learning is introduced to make use of such incompletely classification results for training to improve the performance of classification. This approach can be viewed as a type of computational psychometric method which combine computer science method and theoretic psychometric method.
In this framework, both the misclassification results obtained through DCMs and a student's response pattern are supposed to be conditional distributions of true latent group of this student. Thus, the probability of classification using DCMs giving a response pattern $\mathrm{P}(\mathrm{Y_c}|\mathrm{X})$ can be represented as a multiplying between the probability of true latent group giving a response pattern $\mathrm{P}(\mathrm{Y_c}|\mathrm{T_C})$ and the probability of DCMs

classification giving true latent group $P(T_c|X)$. The aim of DFN is to approximate these two probabilities according to the universal approximation theory. As shown in Figure 1, the proposed DFN framework consisted of following two parts:

1) Firstly, a Co-Train module used DINA and DINO models to classify students' latent group, and the two classification results were used as one target to train the DFN.
2) Considering the misclassification caused by inappropriate DCMs selection, another target, the reconstructed response pattern was added to refine the classification.

The DFN is trained by minimizing a weighted combination of two loss functions: one is the difference between classification 1 & 2 and the DCMs' classification results; the second one is the difference between reconstructed response pattern and observed response pattern.
We conducted both simulated study and real data test to evaluate the performance of our proposed methods. The experimental results showed that firstly, the proposed method is able to achieve similar or better classification result compared with the true general model; secondly the proposed DFN framework is more robust to misspecification of Q-matrix and change of test conditions. The observation proved that our proposed method is more reliable to be applied to real assessment data analysis.

Symposium: Digital Adaptive Learning and Assessment Systems

Authors
Alina A. von Davier (ACTNext by ACT)
Benjamin Deonovic (ACTNext by ACT)
Jingchen Liu (Columbia University)
Viktor Qvarfordt (Sana Labs)
Geoff LaFlair (Duolingo)
Meirav Arieli-Attali (ACTNext by ACT)

Abstract
Adaptive learning and assessment systems have grown significantly in the past few years and have incorporated concepts from assessment, learning, curriculum development, assessment design, and computer science.
In this symposium we present five different perspectives on building adaptive learning & assessment systems: statistical models for diagnostic and logfile data, recommender algorithms for making adaptive & personalized recommendations to learners, the design of tests based on natural language processing (NLP) alignment, and the principled design of tools that allow for the collection of evidence for learning.
The presentations in this symposium provide a broad range of examples of research projects, prototypes, operational capabilities, and products that offer insights into the current landscape of adaptive systems.

An Urn-scheme to Track Accuracy and Response Times in Learning Environments

Author

Benjamin Deonovic, ACTNext by AXCT

An observed response can be conceptualized as an outcome of a match (with a certain duration) between an item and a player. In modelling these observed responses, it is usually assumed that person abilities are not changing -- at least not while making a test. If the skills of players or the difficulty of items possibly change over time (even after every response), this classical assumption of a constant ability fails. This is, for example, the case in learning environments with instructions or feedback.

This presentation proposes an algorithm based on urns that is designed to track changing parameters. The skill of each player and the difficulty of each item is represented with an urn with a certain probability (configuration of balls) and a certain size. This results in binomial distributed random variables with known standard errors. Second, we illustrate the urnings algorithm by analyzing data from a large online adaptive learning environment (Klinkenberg, Straatemeier & Van der Maas, 2011) including both accuracy and response time data. We focus on the (developmental) relation between estimates based on speed and accuracy. Third, we show that this system allows for an adaptive selection of items to the skills of players, without any rating inflation.

An Exploration of Process Data in Computer-based Assessment

Author

Jingchen Liu

In classic tests, item responses are often expressed as univariate categorical variables. Computer-based tests allow us to track students' entire problem-solving processes through log files. In this case, the response to each item is a time-stamped process containing students' detailed behaviors and their interaction with the simulated environment. The key questions are whether and how much more information are contained in the detailed response processes additional to the traditional responses (yes/no or partial credits). Furthermore, we also need to develop methods to systematically extract such information from the process responses that often contain a substantial amount of noise. In this talk, I present several exploratory analyses of process data. These results are applicable to adaptive tests.

Sana Learn: A Personalized Learning Recommender System

Author

Viktor Qvarfordt (Sana Labs)

Sana Learn personalizes educational content based on what students know, how they learn best, and how they forget. Recommending each student's unique and optimal path through a course, helping every student learn faster and stay engaged.

The recommender system gives real-time recommendations by combining offline data analyses on interaction data and real-time algorithms, giving recommendations that are both directly adaptive to student interactions and utilizes features of the entire dataset.

I will present how we model features of educational data, such as the knowledge graph of the learning material, and the behavior and proficiency of each student. We model student learning through predictive models trained on historical interaction data with actual or implicit assessment tests as labels. Depending on the application, we use various deep models: we handle sequential data through recurrent neural networks (RNN) such as long-short term memory (LSTM) networks and use gradient boosted decision trees (GBDT) as ensemble of specific models. We also use classical methods such as logistic regression and Bayesian knowledge tracing, and compute correlations across different data and aggregates.

Furthermore, our algorithm does exploration by presenting alternative sequences of content for students, in order to map out larger parts of the

state space. This allows for causal inference analysis through which we can derive improved recommendations without having to refer to A/B tests.

A Machine-Learned Construct of Language Proficiency and an Adaptive Language Assessment

Author

Geoffrey T. LaFlair (Duolingo)

A central tenet of criterion-referenced testing is that assessments and learning curricula be based on the same set of learning objectives or benchmarks. The development of computer adaptive tests (CATs) and curricula pose similar challenges: they require significant resources in order to develop quality items and quality learning tasks. We report on an adaptive language assessment and language-learning curricula, both of which were built on a set of benchmarks from the Common European Framework of Reference (CEFR) for languages scales. I explain the how the CEFR benchmarks were used to create a machine-learned construct of language ability and how the machine-learned construct was leveraged to create both the CAT and the language learning content.

The machine-learned construct was used to create items and predict the difficulty level of items in a computer adaptive English test. Post-administration item analyses show a strong positive relationship between the machine-learned item parameter estimates and item-response theory estimates based on more than 13,000 test administrations ($\rho = 0.92$). Furthermore, the total test scores from our CAT exhibit high split-half internal consistency (0.91), and show strong positive relationships with other tests of English language proficiency such as TOEFL ($r = 0.73$, n = 1,000) and IELTS ($r = 0.79$, n = 550).

The machine-learned construct was then adapted to other languages (Spanish, French, German, and Italian), and used to create an interactive tool for developing reading and listening materials for specific CEFR-referenced ability levels. This tool takes a text as input, analyzes the linguistic features, and situates it on the CEFR scale. The stories are then made accessible to learners at the appropriate as their language skills develop. The tool's functionality will be demonstrated in the presentation.

Title: Developing a Learning and Assessment System with the expanded ECD Framework; the HERA Showcase

Author

Meirav Arieli (Attali, ACTNext by ACT, Inc.)

Abstract

In this talk I will discuss how to develop a blended learning and assessment system using the expanded ECD for learning (e-ECD; Arieli-Attali, Ward, Thomas, Deonovic & von Davier, 2018). The e-ECD is a framework that incorporates learning goals and learning processes as building blocks at the Student Model, articulating and guiding the design of the Task Model and the learning supports needed in relation to the learning approach, while also specifying the Evidence Model that ensures collecting validity evidence about the construct in focus. I will further illustrate the approach by presenting the HERA prototype, a system to assess and teach scientific thinking that we are currently developing at ACTNext. In the HERA system, we use simulations as a way for students to get familiar with the topic and to learn by doing. Following exploration and collecting data via the simulation, we present students with questions about the data collected. Each question, if answered incorrectly, is followed by three types of help: minimal help of rephrasing the question, medium level help of breaking down the question and providing the first step, and major level of help which provides the wider context and "teaches" the required content needed to be able to answer the question, along with returning to the simulation and the experience of experimenting with it. The three types of help are based on the approach

that each student needs a different type of help, depending on their prior knowledge, ability and experiences. Different scoring and measurement models are considered for these items, from considering only the first response (correct or incorrect) to a partial-credit model. Examples of tasks and the results from a pilot study will be presented and discussed.

Symposium: Development and Validation of three Vertical Scales for Formative Computer Adaptive Testing

Presentation 1:

Mindsteps, Northwestern Switzerland

Authors

Stéphanie Berger (University of Zurich, University of Twente)
Angela J. Verschoor (CITO)
Theo J. H. M. Eggen (University of Twente; CITO)
Urs Moser (University of Zurich)

Abstract

Several studies have pointed to the positive effect of formative feedback on learning and self-regulation. Modern computer technology in combination with algorithms for computer adaptive testing (CAT) can be used as an instrument for providing formative feedback in classrooms on a regular basis. In this presentation, we introduce a tool for formative adaptive assessment targeted to assess students' abilities in several school subjects throughout primary and secondary school in Switzerland. Using the example of mathematics, we describe the development of a vertical scale based on item response theory methods, which represents the heart of this assessment tool. Thanks to the vertical scale, the CAT algorithms can select the most informative items independent of the students' school grade, which is especially relevant for low-ability and high-ability students with one grade. Furthermore, students and teachers can track the students' learning progress not only within one school year but over the course of seven years of compulsory school.

In the first part of the presentation, we elaborate on the technical steps we took for developing the vertical scale for Mathematics. First, we introduce the competence-based curriculum, which served as a content framework for developing the item pool. Second, we illustrate the common-item non-equivalent group design that we developed for assigning items to students from different school grades and, thus, to link items varying greatly in difficulty to one scale. Third, we present results from item calibration and analysis with regard to item fit and parameter invariance between different school grades, and we discuss the criteria we used to exclude misfitting items from the final scale.

In the second part of the presentation, we focus on the validation of the vertical scale for mathematics. In particular, we investigate by means of correlation analyses whether the empirical outcomes of the calibration procedure—i.e., the item difficulty estimates—match to the theoretical, content-related item difficulties reflected by the underlying competence levels of the curriculum. In addition, we explore by means of correlation and multiple regression analyses whether the match between the empirical and the content-related item difficulties differ for items related to different curriculum cycles, domains or competences.

Presentation 2: *Nationale test og folkeskolens digitale prøver, Denmark*

Author
Jakob Wandall (NordicMetrics, University of Aarhus)

Abstract
In 2005, a computer adaptive test system for national tests in Denmark was commissioned via a EU-tender. It was described in the tender that the provider should deliver 12 adaptive testes for the Danish public school (grade 2-8). Every test was targeted a grade level and designed to measure three constructs (called profile areas) in one of the following a subjects: Danish/reading (grade 2, 4, 6, 8), Mathematics (grade 3, 6), English (grade 7), Physics, Biology, Geography (all grade 8) and Danish as a second language (grade 5, 7). In 2006 an act was adapted in the Danish parliament that made the first 10 of these tests mandatory at the grade level they were targeted. In the contract with the provider it was described that there should be created 12 x 3 item banks (each containing at least 150 items with a large variety in difficulties) and each of them should fit Rasch Model. All items were pretested on 700 students (most on the grade level and some on grade levels above and below). For Danish/reading and Mathematics, the items should be tested with anchor items in a way so that the scales in the same subjects could be linked vertically later on. It took far longer than planned to establish the item banks and the vertical linking was never done – the tests were launched full scale in 2010. In 2018 two more tests (Math in 8, and English in 4'th grade) were added to the system.
The presentation consists of three parts:
1.A short description of the Danish National CAT-system and the background for the design. The tests are designed to be used for formative evaluation, they are low stakes and each student can take each test 3 times (two tests are voluntary) on the grade level below to the grade level above.
2.A description of the basic scaling (by Rasch-modelling) and development of a vertical scaling of the reading and math tests by equating the total Rasch scores based on CCT techniques.
3.A demonstration of the visualizations of the results for formative purposes and pedagogical use by combining the information in the vertical scales whit the advanced analytical features in the test system.

Presentation 3: *The RIT scales by NWEA, USA*

Author
Gage Kingsbury (Psychometric Consulting)

Abstract
How do we show that a measurement scale meets vertical, interval scaling assumptions within the context of an operational adaptive test? Much research has been done on procedures to create such measurement scales (Tong and Kolen, 2007.) Some research has been done looking at existing scales and the impact their characteristics may have (Briggs and Betebenner, 2009 clarifies the issues.) Little research has been done concerning vertical scales used in with adaptive tests. This study begins to examine the performance of a measurement scale when it is used for adaptive testing across grades.
This live data study investigates the NWEA RIT scales used in K-12 education. These scales were designed to measure growth from one grade to another, which requires a vertical scale, a large pool of items, and an adaptive testing design that allows content to be controlled according to

the grade that a student is in. This study examines whether each item performs similarly in each grade in which it is administered. If this item fit varies from grade to grade, it is evidence that the scale does not meet the assumptions of a vertical scale. If the fit for all items is consistent across grades, it serves as one piece of evidence for the use of the vertical scale.
This real-data study examined 20,000 mathematics tests administered to students in grades 3 to 9. Each test was an adaptive test consisting of 50 precalibrated items drawn from a pool of over 2000 items, used across grades. For each item administered to 20 or more students in more than one grade, several misfit indices were calculated in each grade, and were examined conditional on estimated achievement.
The results indicated that 99.7% of the items administered in multiple grades had an acceptable level of fit in each grade. This trend was consistent when students were categorized into different trait levels. While these overall findings support the quality of the vertical scale, the items that do not fit equivalently across grades warrant additional study. These may be instances that show that items have different properties as measurement tools. This may improve our understanding of items, which may improve measurement of student achievement.

Symposium: From Adaptive Testing to Artificial Intelligence and Smart Learning

Authors
Hua-Hua Chang (Purdue University)
Nathan Thompson, (Assessment Systems Corporation)
Shengyu Jiang (University of Minnesota)
Chun Wang University of Washington)
Yinhan Chen (University of Nevada, Reno)
Shiyu Wang (University of Georgia)
Susu Zhang (Columbia University)
Xueying Tang (Columbia University)
Zhi Wang (Columbia University)
Jingchen Liu (Columbia University)
Zhiliang Ying (Columbia University)

Abstract
The theory and methods of Computerized Adaptive Testing (CAT) have been well advanced during the last 40 years. The rapid developments in technology have made large-scale CAT implementations easier than ever before. However, CAT methods have not been well acknowledged or even known by many researchers in Artificial Intelligence (AI). Today we would like to highlight our discussion whether CAT can help AI in educational research, in particular, how to better design “deep learning” and “neural network” to archive many attractive missions, such as smart testing and smart learning. We will show how CAT can be utilized to build a tailored assessment for each individual in the big data era. Our goal is to build many reliable, and also affordable, web-based diagnostic tools for schools to automatically classify students' mastery levels

for any given set of cognitive skills that students need to master. In addition, we will show how the tools can be employed to support individualized learning on a mass scale.

Symposium: From Adaptive Testing to Artificial Intelligence and Smart Learning

Hua-Hua Chang, Purdue University
Nathan Thompson, Assessment Systems Corporation

Abstract

The theory and methods of Computerized Adaptive Testing (CAT) have been well advanced during the last 40 years. The rapid developments in technology have made large-scale CAT implementations easier than ever before. However, CAT methods have not been well acknowledged or even known by many researchers in Artificial Intelligence (AI). Today we would like to highlight our discussion whether CAT can help AI in educational research, in particular, how to better design "deep learning" and "neural network" to archive many attractive missions, such as smart testing and smart learning. We will show how CAT can be utilized to build a tailored assessment for each individual in the big data era. Our goal is to build many reliable, and also affordable, web-based diagnostic tools for schools to automatically classify students' mastery levels for any given set of cognitive skills that students need to master. In addition, we will show how the tools can be employed to support individualized learning on a mass scale.

*1.**AI and Machine Learning in Psychometrics? Old news.***

Nathan Thompson, Assessment Systems Corporation

Abstract

In the past decade, terms like machine learning, artificial intelligence, and data science are becoming greater buzzwords as computing power, APIs, and the massively increased availability of data enable intriguing new technologies like self-driving cars. However, we've been using methodologies like AI and machine learning in psychometrics for decades. So much of the hype is just hype. And yet, we still massively underutilize this technology. Adaptive testing has been around for 40 years but remains used by only a tiny percentage of programs. This presentation will discuss some of the ways that AI and ML are currently used to improve assessment, and where new technologies afford incredible opportunities. We frequently hear how half of white-collar work might be replaced in the next two decades; where will this occur in the assessment industry? Because, without a doubt, it will happen.

*2.**On-The-Fly Parameter Estimation Based on Item Response Theory in Adaptive Learning Systems***

Shengyu Jiang, University of Minnesota
Chun Wang, University of Washington

Abstract

Item response theory (IRT) has seen many successful applications in computerized adaptive testing (CAT), yet its use in a learning scenario is limited. In an adaptive learning system, the learning materials are matched to the level of the student to facilitate the learning process. Compared to CAT, the adaptive learning system requires a much larger calibrated item bank as well as a model that allows dynamic changes of learners' abilities. We applied a recently developed Bayesian algorithm to the learning environment. The algorithm can capture the change of student ability, while at the same time calibrate the parameters of items with only part of the item parameters known. It is demonstrated

through a simulation study that that this algorithm is well suited to building adaptive learning systems due to its fast computation and cost-efficient item calibration.

*3.**Automated Attribute Hierarchy Detection with Application to Adaptive Learning***

Shiyu Wang, University of Georgia
Yinhan Chen, University of Nevada, Reno

Abstract

Adaptive learning or personalized learning is a relatively new technology assisted learning framework that refers to "instruction in which the pace of learning and the instructional approach are optimized for the needs of each learner". Attribute hierarchy or skill hierarchy, that is the mastery of one attribute is a prerequisite to the mastery of another one, can be used to facilitate in designing an adaptive learning system. The attribute hierarchy is commonly established by cognitive experts in a specific content domain. In this research, we explore the possibility of using Bayesian Graphical Model, Markov random field and Neural Network to automatically detect the attribute hierarchy from the learning assessment data. The proposed techniques are applied to several learning data sets to construct attribute hierarchy for different subject domains. We finally present an adaptive learning design by using the detected attribute hierarchy and a learning model that can quantify the instructional effectiveness of different learning materials.

*4.**Understanding Interactive Items' Characteristics by Deep Learning-based Process Data Analysis***

Susu Zhang, Columbia University
Xueying Tang, Columbia University
Zhi Wang, Columbia University
Jingchen Liu, Columbia University
Zhiliang Ying, Columbia University

Abstract

Computer-based testing has enabled the administration of interactive testing items, which allows for the measurement of traits that are hard to assess with traditional, non-interactive items. The sequence of actions that an examinee performs on an interactive item, as well as the amount of time elapsed during and between actions, document the respondents' problem-solving process. They can hence entail information about the respondent's characteristics that cannot be recovered solely from the final responses. The current study explores how much we could learn about individual characteristics based on the log data and the corresponding interevent times from an interactive item. We use a recurrent neural net (RNN) to predict various characteristics of the respondents, including different cognitive scale scores and demographics. Besides evaluating the prediction power of each item's log and timing data, we further seek to decompose these items' prediction power by (1) interpreting the RNN-generated features, (2) identifying key actions/subsequences of actions, and (3) examining the improvement in prediction accuracy by adding timing information of specific actions (on top of action sequences without timing information). By studying these systemic behavioral differences related to underlying individual characteristics, test developers can gain an additional perspective on the evaluation and design of interactive items.

Relaxing Constraints for Better Measurement Precision during a CAT

Authors
Jonghwan Lee
Sylvia Scheuring
Sukkeun Im
Jungnam Kim
Christy Schneider

Abstract
During a computerized adaptive test (CAT) administration, items are selected to maximize the measurement precision at each individual student's estimated level of ability. When the adaptive engine selects items, it considers multiple factors such as maximizing item information, meeting the test specifications, and controlling item exposure, all of which rely on the quality of the item pools. Adaptive item pools should have sufficient numbers of items with varied item difficulties across blueprint categories so the adaptive engine can administer items at the appropriate difficulty level given each individual student's ability level. However, for heavily constrained CATs such as those for a summative statewide assessment, it can be very difficult to have enough items in the pool that can cover all the blueprint requirements in a real testing environment while maintaining measurement precision for each student. Therefore, administering a CAT may require a certain level of compromise and flexibility between measurement precision and content coverage requirements.
This simulation study compares blueprints constrained at the strand level and indicator level on measurement precision at various ability levels estimated using an item pool with a limited number of high-difficulty items. Bias, RMSE, item exposure rate analyses, SEM, and reliability results will be overviewed. Results indicate that both sets of constraints recovered student ability fairly well and were similar to each other. Relaxing the blueprint constraints to allow focus on the strand level of the content standards while also requiring at least one item at each indicator increased the measurement precision at various levels of student ability while also allowing test events to meet common criteria found in alignment studies.

On the impact of measurement error in computerized adaptive testing

Author
Angela Verschoor (Cito)

Abstract
One of the main advantages of computerized adaptive testing (CAT) is an enhanced measurement efficiency. In order to reach this, items are usually selected according to the amount of information they provide in, for example, a specific region on an IRT-scale. This amount of

information is calculated using item parameters that have been estimated, while usually uncertainty in these parameters is not taken into account in the item selection process. Maximizing Fisher information tends to favor items with positive estimation errors in the discrimination parameter and negative estimation errors in the guessing parameter. This is also referred to as capitalization on chance. Not taking the uncertainty into account might be a serious threat to both the validity and viability of CAT. Previous research showed quite an effect on the resulting test information function (TIF), whereby the TIF was overestimated by as much as 20%. For fixed-length CATs, the measurement accuracy might therefore be overestimated, while stopping criteria based on this accuracy may result in unduly short test sessions.
In this presentation, a robust method for CAT is presented as a method that accounts for uncertainty in the item parameters. The method consists of two phases. In the first phase, the parameters of the item pool are transformed into robust parameters, while in the second phase the item selection takes place in the usual way using those robust parameters. In a simulation study, the impact of measurement error are shown, as well as the effects of robust CAT in compensating for this measurement error. The overestimation of the TIF turned out to be very small, usually well below 1%. Some theoretical considerations are shared. Finally, the implications are discussed.

Enhancing Effectiveness of CAT Simulation with the Shadow-Test Approach in RSCAT

Authors
Bingnan Jiang (ACT, Inc.)
John Whitmer (ACT, Inc.)

Abstract
Computerized adaptive testing (CAT) integrates modern computer technology with measurement theory to improve the efficiency of assessment for both the testing provider and the learner by selecting the best items to administer to measure the ability of an individual examinee. High-quality assessment in CAT relies on well-managed test content to represent the examinee's skills and knowledge. In addition, statistical models need to accommodate differences between items, e.g., some items are easier than others. A fundamental dilemma in CAT is to administer optimal items sequentially while meeting content specifications. In this presentation, a recently-released open source R package that uses a shadow-test approach to computerized adaptive testing (RSCAT) is described to address this dilemma for CAT research and simulation. The shadow-test approach (van der Linden, Linear models for optimal test design, 2005) solves the dilemma by dynamically assembling entire shadow tests as a part of selecting items throughout the testing process. Shadow-test CAT has many advantages, including full coverage of the test blueprint, separation of test specifications from CAT algorithms for easily modifiable configurations, and supporting flexible and reliable delivery options. There exist open-source packages for CAT research and simulation using the shadow-test approach, e.g., mirtCAT and xxIRT. However, they have limitations in terms of runtime performance, supported constraint types, and usability. RSCAT is different from existing CAT packages that use the shadow-test approach in three aspects. First, as the lightweight version of Echo-Adapt, the commercial CAT engine used by ACT, Inc., RSCAT guarantees the correctness and high efficiency of CAT simulation. CAT models and algorithms in RSCAT are implemented in Java and wrapped with R APIs for external calls by other R programs. Second, RSCAT expands the shadow-test approach in many aspects, e.g., modeling the shadow-test mixed integer programming (MIP) in a concise and well-structured way, extending the types of test specification

constraints through designed syntax, and supporting various MIP solvers for shadow test assembly. RSCAT allows users to bring their own commercial MIP solvers, e.g., CPLEX and Xpress, to maximize the efficiency of solving shadow-test assembly MIPs in a large-scale simulation or simply use open-source solvers for quick research validation. Third, a Shiny app is implemented to assist users with a graphical interface to configure their CAT simulations. With RSCAT, users can effortlessly configure test blueprints and CAT algorithms even though they have no background in R/Java programming. In the presentation, the fundamental models and algorithms for shadow-test CAT used in RSCAT will be discussed. The software architecture of RSCAT will be described with highlights on key implementation technologies. We will also demonstrate the process of configuring CAT and running simulations through the Shiny app and R APIs. The simulation results and performance metrics demonstrate the efficacy of the design and implementation of RSCAT. In conclusion, RSCAT is a powerful and open-source R package that enables shadow-test CAT with enhanced effectiveness, efficiency, and usability.

Penalization to Item Selection Inadequacy in a CAT

Authors
Chen Li
Michael Chajewski

Abstract
In traditional computerized adaptive test (CAT), items are selected to achieve minimum measurement error of an examinee's ability, while satisfying a variety of targeted assessment configurations. Gonulates (2015) proposed the item pool utilization index (IPUI) to evaluate the efficiency of an item pool for adaptive selection. The IPUI, which compares the observed item information and the maximum possible information, is summarized across items and examinees. Because of this, the IPUI is insufficient in assessing item selection at the individual assessment sequence level. At a sequence level, it is expected that item difficulties for administered items are not consistently higher or lower than the ability estimates along consecutive positions. In a practice CAT, an item pool often supports multiple CATs per examinee with non-repetitive items. With such item pool, the IPUI may indicate that the item selection is satisfying, yet some sequences, as later attempts of an examinee, may (a) consistently administer items with item difficulty higher or lower than the real-time ability estimates on consecutive positions and/or (b) have large discrepancy between item difficulties and real-time ability estimates, as a result of insufficient items in one or several specific content categories. In addition to IPUI, it is important to monitor item selection efficiency at sequence level to ensure effective and consistent CAT experiences for examinees.
This study proposes a modification to the IPUI. The proposed approach penalizes selecting an item where the difference between the ability estimate and item difficulty has the same sign as that of the previous position. By assigning a within item selection difficulty interval derived weight to the quotient of observed to maximum possible item information, insufficiencies in item adequacy (including both signage and ability-difficulty discrepancy) are magnified to reflect the reduction in efficiency. The proposed moderation is as follows.

$$IPUI_{seq} = \frac{\sum_{k=1}^{K} \frac{I_k[\hat{\theta}_{k-1}]}{I_m[\hat{\theta}_{k-1}]} * w_k}{\sum_{k=1}^{K} w_k}, where \begin{cases} N_{\pm_k} = 1, & w_k = \frac{R_k}{R_k - |\theta_{k-1} - b_k|} \\ N_{\pm_k} = 0, & w_k = 1 \end{cases} \quad (1)$$

In Equation 1, the modified IPUI for a sequence is the weighted mean of the quotient of observed to maximum possible item information, presented as $\frac{I_k[\hat{\theta}_{k-1}]}{I_m[\hat{\theta}_{k-1}]}$, across all K items in a CAT sequence. The weight w_k is conditional on whether the sign of difference between item difficulty and ability estimate is the same with that on the previous position. If the signs are different, $N_{\pm_k} = 0$, assign a weight of 1 to the information quotient. Whereas $N_{\pm_k} = 1$ indicates the same sign, the information quotient is outweighed by assigning the a weight of the ratio of the item difficulty selection range on position k, denoted by R_k, to the difference between the item selection range and the absolute discrepancy between item difficulty on position k and ability estimates on position $k - 1$. The larger the absolute difference between item difficulty and ability estimates, the severer the penalization.

The modified IPUI is evaluated using data from an operational criterion referenced variable length CAT. The performance of the IPUI and its proposed modified version are evaluated and compared based on their effectiveness of detecting the consistency in item selection signage and the influence of asymmetric selection on the substantive nature of the assessment.

An Investigation of Item Selection Criteria for Passage-based Items

Author
Catherine Mintz
Deborah Harris Ph.D.
Aaron McVay

Abstract

As our testing culture veers toward adaptive technology as its chosen modality, it is increasingly important to ensure tests maintain sufficient measurement properties. Much work has been done already, notably on the comparability of CAT scores to fixed-format test scores (e.g., Thompson & Davey, 1999). However, it is not only test modality that needs to be vetted; the impact of item format must also be considered. Passage-based items, generally scored or measured using polytomous modeling procedures, constitute a common item format with complex measurement possibilities. This study focuses on passage-based item selection under different item selection methods and when passage effects are not accounted for in statistical models.

Passage-based items were calibrated under the 3PL. Within each testlet, the average item difficulty and average MI were computed. Testlets were selected if their average item difficulty was closest to the simulee's $\hat{\theta}$ (average CSV) or if they provided maximum information for a simulee's current $\hat{\theta}$ (average MI). In two conditions, items were calibrated under the 3PL and administered

individually using either CSV or MI; these conditions served as baseline procedures. Lastly, a condition was run where testlets were selected if one of the items within the testlet had an estimated difficulty closest to $\hat{\theta}$ (CSV with Testlets).Though this test assesses an overall science domain, the items correspond to science subdomains (i.e., earth science, life science, and physical science). We did not control for these subdomains. However, we reviewed administered CATs to retroactively investigate how content balancing was affected by item selection method. We also retroactively examined exposure rates to determine how item selection procedure affected exposure.

The five CATs produced similar and encouraging results. Overall, $\hat{\theta}$ tended to highly correlate with θ. MI produced the highest correlation and CSV with Testlets produced the lowest. Corroborating results from Koch and Dodd (1995), who found that CSV and MI produced similar $\hat{\theta}$s, we found that CSV and MI produced similar correlations between $\hat{\theta}$ and θ. However, these methods appear to administer tests with different items and different content proportions. Overall, θ tended to be slightly overestimated. Bias was small and nearly identical in each condition. Nonetheless, MI clearly produced the most precise results and CSV with Testlets produced the most imprecise. In all conditions, higher levels of θ were estimated more precisely than lower levels. The content proportions in the CSV condition most closely resembled those from the item pool. Content in the CSV with Testlets condition was remarkably unbalanced: most items administered were life science items; few physical science items were administered. In the averaged conditions, content was roughly balanced. No clear pattern exists to indicate what type of overall selection method maintains content balance best.

A concern is that passage-based items not administered in sequence immediately become enemies. Using an unaveraged item selection method for passage-based items automatically creates item enemies. From a test development perspective, it is crucial that passage-based items be administered together. If content balancing is not a concern, either of the averaged methods could be used.

Validity Evidence from a Formative CAT Educational Mathematics Assessment

Author
James B. Olsen (Psychometrician, Renaissance)

Abstract

Validity evidence is presented for unidimensionality and parameter estimation invariance with a formative, Rasch CAT mathematics assessment. Domain score unidimensionality was evaluated with Exploratory Factor Analysis and Confirmatory Factor Analysis using 201,000 assessments sampled across four grade bands: K-2, 3-5, 6-8, and 9-12. CFA was conducted overall and within grade bands. CFA investigated the unidimensional structural equation model shown in Figure 1. Two CFA models were tested: four domains and three domains.

Table 1 summarizes the CFA analysis by grade band including multiple fit statistics. Eigenvalues > 1.0 were retained. First factor loadings per domain were all above 0.80.The Root Mean Squared Error Approximation and Standardized Root Mean Residual values < 0.08 show evidence of unidimensionality. The Comparative Fit Index, Goodness Fit Index and Normed Fit Index values are near 1.00, providing strong unidimensionality evidence. Table 2 presents CFA factor loadings for subscore domains. Results show high factor loadings within and across grade bands for three and four domains. CFA factor loadings between 0.78 to 0.93 show consistent structure within domains across grade bands, and within grade bands across content domains. Table 3 summarizes EFA analyses across grade bands for the CAT. Factor loadings for seven different EFA analyses range between 0.90 and 0.97.

Evidence of parameter estimation invariance was investigated across years, months, seasons, with multiple calibration periods of 2-6 months. The empirical analysis included 494 mathematics items, grades 1-10, from an online calibration environment fielded for 2+ school years with 4.6 million item responses. Effects on calibrated item difficulty were investigated from calibration periods of different durations. Rasch parameter estimates from the full data set were the “gold standard” against which estimates from shorter calibration periods were compared. Five criterion statistics included: Root Mean Squared Difference(RMSD), Mean Absolute Difference(MAD), Mean Difference(MD), Pearson correlations, and effect sizes. Thresholds of 0.50 logits were established for RMSD and MAD, 0.20 logits for MD, and 0.20 for effect sizes. The Winter calibration season exhibited descriptive statistics comparable to the “gold standard”. Descriptive statistics from Fall and Spring calibration seasons were slightly lower than the “gold standard”, but within acceptable ranges. Separate Fall, Winter and Spring calibration seasons each showed RMSD and MAD values less than the 0.50 threshold criterion, MD values < 0.20, very high correlations with the “gold standard”, and small effect sizes.

Monthly calibrations showed RMSD and MAD values less than the threshold of 0.50 logits with exceptions of RMSD for December (0.5235) and June (0.5835). MD results were less than 0.20 for all calibration periods. Monthly correlations ranged from 0.965 to 0.995, median 0.987 ; multiple month correlations were above 0.990. Effect sizes were very small, less than 0.05.

Parameter estimation variability was reduced by lengthening calibration periods. Scatterplots exhibited greater variability for monthly calibration periods than multiple bi-month periods, and sequentially less variability for calibration periods of 3-6 months duration. Stable parameter estimation results when each item is administered to 1000 examinees, fielded for contiguous two-month periods, and administered both on grade level and one grade level above.

Interval Validation Method: Achievement Level Setting Based on Large Item Pools Used in Computerized Adaptive Testing

Author
William R. Insko, Jr., Ph.D. (Houghton Mifflin Harcourt)

Abstract
The Interval Validation Method or IV Method (Insko, 2018; Insko & Murphy, 2016; Insko & Murphy, 2017) for setting achievement level standards is specifically designed for assessments with large item pools, especially computerized adaptive tests. The method takes advantage of the fact that large pools of calibrated items that have been ordered by difficulty have multiple, unique sequences of items with very similar difficulties, within several tenths of a logit. The IV Method breaks item pools into intervals so experts can be guided to *locate*, *verify*, and *endorse* the interval that best aligns with the performance level descriptors. Endorsing the interval sets the cut score.
The proposed paper will include a detailed description of the steps for implementing the IV Method, including those aspects of the IV Method that make it unique from other standard setting methods (Cizek & Bunch, 2007). In addition, results from an operational achievement level setting workshop will be reviewed, along with comparisons of impact data from several SBAC school districts. Finally, research related to item plats specifically designed for the procedure will be reviewed.
The item plat research is particularly important because item plats lists items from the item pool in order from easiest to most difficult based upon IRT item difficulty (for example, Linacre, 2006; Rasch, 1960). The “Interval” and “Exemplar” columns included in the item plats define the intervals to be used during standard setting. Research on the optimal number of items to include in an IV Method interval and strategies for reducing item pool size to make standard settings more manageable have both been systematically investigated through simulation studies. In addition, issues related to balancing domain coverage within intervals have also been studied. The paper will review current recommendations for selecting an optimal interval length, a strategy for systematically reducing item pools, and strategies for balancing domain coverage within intervals.

Selecting off-grade items appropriately in a vertically-scaled adaptive test

Authors
Shuqin Tao (Director of Psychometrics, American Institutes for Research)
Luciana Cançado (Data Scientist, Curriculum Associates)
Brett Morrow (Associate Director of Assessment, Curriculum Associates)

Abstract

It is quite common to have a computer adaptive test built on a vertically-scaled item pool, in which students may see on-grade items as well as items from below- or above-grade levels, as a result of the adaptive algorithm searching for items best targeted to the estimated student ability level. However, a question that lies at the core of this practice remains largely unanswered: should the item selection algorithm be agnostic of item grade? Put in more concrete terms, suppose the item selection algorithm identifies two items with the same difficulty for a 4th grader, and further suppose we know one item is from 4th grade and the other from N grades above or below, should we randomly choose between these two? The answer likely is: "it depends". Some assessment programs may have designed some general rules regarding the use of off-grade items based on the best content knowledge. However, it is important to check these content-based rules using empirical evidence and refine them if warranted. This is the primary motif of the present study.

Measurement invariance is the foundation of adaptive testing, which ensures that students receive comparable scores when given different sets of items. For measurement invariance to hold in the context of selecting items agnostic of item grade from a vertically-scaled item pool, it is important to ensure items do not function differentially with respect to grade level. To the extent that an item is differentially easy or difficult for students of the same ability from different grades, it exhibits differential item functioning (DIF) with respect to grade level. To the extent items show grade-level DIF, the use of off-grade items may violate the measurement invariance principle, a consequence of which is that students of the same ability may receive different ability estimates depending on the makeup of the items they received regarding item grade.

This study capitalized on the data that came from a vertically scaled computer adaptive assessment for students in kindergarten through grade 8. We conducted the DIF analysis using a procedure embedded in the WINSTEPS program, in which K-8 concurrent calibration was conducted and the DIF analysis was conducted by selecting the student grade as the person variable. DIF contrast and Mantel-Haenszel (MH) DIF statistics produced in WINSTEPS output were examined to answer these research questions:

(1) To what extent do items exhibit grade-level DIF and in what pattern?

a. How do reading and math compare?

b. How is grade-level DIF related to the discrepancy between item grade and student grade?

(2) What might be some underlying causes for grade-level DIF? How do they differ by subject, domain and grade?

Figure 1 contained some preliminary findings for math, which would likely shed light on the scope, pattern, and nature of grade-level DIF and help us better understand the appropriateness of using off-grade items from both psychometric and substantive perspectives. This understanding will help us refine the item selection algorithm regarding the use of off-grade items in a vertically-scaled adaptive test.

Consideration of Item Position Effects in CAT with the Continuous Calibration Strategy

Authors

Andreas Frey (Goethe University Frankfurt, Germany; Centre for Educational Measurement (CEMO) at the University of Oslo, Norway)

Aron Fink (Goethe University Frankfurt, Germany)

Christian Spoden (German Institute for Adult Education - Leibniz Centre for Lifelong Learning Bonn, Germany)

Abstract
In computerized adaptive testing (CAT), knowledge about the item parameters of the test items in the pool is required to select the next item. These item parameters are estimated based on item responses collected in a calibration study using an item response theory (IRT) model. In several potential application areas of CAT, constructing large numbers of items prior to the test's initial use, and/or carrying out a calibration study with a large sample is not feasible. Correspondingly, for applications such as standardized written exams, psychological tests used in personnel selection, for clinical diagnosis testing or in research, CAT is typically not used, even though it would be advantageous here too. The recently proposed continuous calibration strategy extends CAT's application range to such application areas. It is useful if a test is administered at many points in time with a fixed length. The continuous calibration strategy is applicable when setting up a new adaptive test or when converting a linear test to a computerized adaptive test. The basic ideas of the strategy are (a) item calibration oriented to the time and capacity available for test development, (b) utilizing item responses across recurring assessments for item calibration purposes, (c) maintaining the reporting scale over time, (d) continuously increasing the adaptivity of the test during its operational use, while (e) reaching a high level of robustness by checking for item parameter drift and controlling for item position effects. To control for item position effects, the positions in which items are presented in the test are balanced across test cycles (repetitions of the test) on the level of blocks of items. In the presentation, I will describe the key elements of the new continuous calibration strategy and present results from a comprehensive Monte-Carlo simulation study examining its capability to accommodate for item position effects. This simulation study is based on a factorial design with the between-factors IRT model (1PL, 2PL), sample size per test cycle (100, 300, 500), item position effect (none, small, medium, large), number of item blocks (3, 6, 12), and the within-factor test cycle (1, …, 11). The dependent measures are Bias and MSE of the ability estimates. The R-based simulation is currently running but will be completely finished and analyzed prior to the conference. The results will provide detailed insights into the capability of the continuous calibration strategy to avoid biased and/or imprecise ability estimates due to item position effects in CAT. The results will be discussed and software for applying this new method will be mentioned.

Enemy Item Identification Using Data Mining Methods

Author
J. B. Weir

Abstract
Enemy items are any two items that should not appear on the same test form. These items may address the same material, or one may provide clues about the answer to another. Most often, enemy item pairs are identified before forms are published; subject matter experts (SMEs) manually review forms for enemy pairs, a process that can be both cognitively taxing and expensive. Indeed, in a CAT context, where a form is the entire pool of items, the review of all item interactions can be especially taxing.
Some researchers have suggested statistical approaches for identifying enemy item pairs; for instance, response data might show violations of local independence caused by clueing. One drawback, however, is that these are *post hoc* tests: the forms must have been administered to a sufficient number of examinees prior to determining enemy relationships.

This study proposes a method of identifying enemy item pairs that capitalizes on two data mining approaches: 1) latent Dirichlet allocation (LDA), an unsupervised topic model that comes from a natural language processing (NLP) context, and 2) a random forest classifier, a supervised ensemble learning algorithm. Output from the LDA model is used to calculate the Jensen-Shannon distance (JSD) between items, which is a measure of topic similarity. Random forests are trained with and without the JSD, as well as several other item-level metadata variables. Item pairs are scored using the resulting random forest classifiers, and SMEs evaluate the output. The random forest classifier is then retrained using input from the SMEs.
This study of 7,205 operational items suggests that random forest models can be useful in the identification of enemy item pairs; furthermore, information derived from the LDA topic model improves the performance of the random forest classifier. Finally, integrating feedback on classification from 29 subject matter experts further improves the performance of follow-up random forest models.

CAT system utilizing automated procedure for detecting enemy items based on NLP

Author
Haruhiko Mitsunaga

Abstract
Computer Adaptive Testing (CAT) technology plays an important role in assessment of L2 learners, especially reading skills. Most reading comprehension tests contain items which measure how test takers use proper vocabulary. Such test forms may include pairs of items which lend themselves to be easily answered because they share related topic or knowledge. These items are called 'enemy items' and should be eliminated from any item series of CAT.
For administrating a vocabulary test using CAT, a large number of items need to be stored in an item bank. However, as more items are added, the number of item pairs also increase. For this reason, the item bank requires close monitoring of the enemy items it may store, and some automated system should be implemented to detect potential enemy item pairs.
Proposed automated approach
This presentation proposes an automated system for detecting enemy items. This system uses a natural language processing (NLP) approach, based on machine learning technology which vectorizes words or series of terms in an item bank so that the similarity of meaning can be compared on a quantitative scale. Cosine similarity index are frequently used to represent distance of meaning on the scale, and if the word similarity between the correct answer of a focal item and another item is extremely high, these two items are detected as potentially enemy items. The main idea of this approach pertains to the technology of obtaining word vectors, such as Word2Vec. However, it requires a trial and error approach to interpret the results properly. Through a learning process, the proper corpus which covers the entire vocabulary space is needed and the proper value of parameters, such as window size, and number of dimensions of vector must be supplied. Two models of representing word similarity are proposed: CBOW and Skip-gram, these two methods returning different types of vector. Existing studies have estimated word similarity under only a single condition, and their methods of summarizing results are yet to be proposed.

Aim of this study
In this study, a procedure for aggregating vector estimates under various models and parameter conditions is proposed. This process enables us to represent similarity index of item pairs in an item bank using vector estimates for diverse conditions. An examination of monitors were conducted to test for the validity of this method, and a CAT system utilizing this method is proposed.
Result and discussion
According to the results of the monitor exam, item difficulty index increased when the items which had been detected as enemy items were presented. However, even when the most proper corpus was provided to execute machine learning process, several pairs of English vocabulary items which were of different topics were designated as enemy items. To reduce probability of such false alarms, a special corpus with more specific area of textbooks in English as a Lingua Franca may be useful.

Trends in Rapid Guessing Behavior Within a Low-Stakes K-12 Computer Adaptive Test

Author
Audra Kosh

Abstract
The proliferation of computer-based testing has sparked research related to K-12 students' rapid guessing behavior whereby examinees quickly click through items without exhibiting effort (Wise & Kong, 2005). Such behavior threatens the validity of test score inferences because random guesses introduce construct-irrelevant variance in students' scores, a threat that may be heightened on low-stakes tests because students' have fewer reasons to try their hardest. For this reason, it is important for testing programs to understand trends in rapid guessing behavior as related to characteristics of students and the test. The purpose of this study is to examine these trends in a diagnostic K-12 computer-adaptive test by answering the research question: how does the frequency of rapid guessing behavior vary by student grade level, content area (Mathematics, Reading, Language Arts), season of test administration (Fall, Winter, Spring), item type (multiple-choice and technology-enhanced), and item delivery sequence (e.g. first item, second item, etc.)?
Data were obtained from variable-length computer-adaptive tests in each of Mathematics, Reading, and Language Arts administered to students in grades K-12 during three seasonal testing windows in the 2017-2018 school year. On average, students completed 45 items with a maximum of 65 items. In preliminary analyses, items were classified as a rapid guess if the student responded to the item in less than three seconds (Kong, Wise, & Bhola, 2007).
Preliminary results from 192,702 Mathematics test sessions (approximately 15,000-25,000 test sessions per grade in K-8 and between 300 and 3,000 test sessions per grade in grades 9-12) indicated that students in higher grade levels exhibited more rapid guessing than lower grade levels, and rapid guessing was more frequent as the school year progressed (Figure 1). Rapid guessing was also more frequent for items administered later in the test session: 0.018% of items administered first were rapidly guessed as compared to 33.6% of items administered in position 65 (i.e., the last possible item administered). However, rapid guessing behavior was minimal overall: 93.3% of test sessions exhibited rapid guessing on less than 5% of administered items within the test session (Table 1). Further analyses, to be completed before IACAT, will

expand upon the preliminary Mathematics results presented here and will include similar results for Reading and Language Arts in addition to considering alternative criteria for identifying rapid guesses (e.g., Wise & Ma, 2012), such as adjusting the three-second threshold upward for technology-enhanced items and for kindergarten students still developing fine-motor skills applicable to computer mouse use.
Results of the study will provide other low-stakes K-12 testing programs with an understanding of rapid guessing behavior in order to modify test administration procedures in ways that may reduce rapid guessing of future test takers. For example, testing programs may consider prompting students with encouraging messages after detecting rapid guessing behavior or may warn teachers or administrators of potentially invalid scores in cases of serial rapid guessing, and these prompts could conditionally appear based on characteristics associated with more frequent rapid guessing (e.g., spring testing window, high school grade levels).

A Comparison of Three Approaches That Permit Answer Review and Revision in Computerized Adaptive Testing

Authors
Yanyan Fu
K. Chris Han

Abstract
In the sphere of educational test measurement, computerized adaptive testing (CAT) has become increasingly popular because it provides a better estimate of test-taker ability than traditional paper and pencil testing. The typical CAT algorithm, however, does not allow examinees to revise their answers to previous test items because each automated selection of a new item depends on the test taker's responses to prior items. Such a restriction may produce test anxiety among some examinees, which can make them "…feel at a disadvantage when they cannot review and alter their response" (Lunz, Bergstrom & Wright, 1992, p. 34). Stocking (1996) also suggests that not allowing examinees to review and revise answers can threaten "test fairness and accuracy." (p. 8)
Several creative approaches have been proposed to allow test takers to review and revise items in CAT-administered examinations. The approach proposed by Lunz and colleagues (1992) allows examinees to review items after they have answered all questions. The validity of this approach in CAT is threatened, however if Wainer's (1993) strategy is used. Specifically, if an examinee intentionally answered all questions incorrectly before reviewing them, he or she would receive only easy questions. The estimation of ability thus would be inflated.
Because Stocking's research (1996) showed that the percentage of items that test takers revised was as low as 3% in linear testing, she proposed a modified end-of-test review (MEOTR). In Stocking's approach, the number of items a test taker can revise is limited to a fixed quantity to reflect the linear testing condition and counter Wainer's strategy. More recently, Han (2013) proposed an item pocket (IP) approach, which allows examinees to skip up to a fixed number of test items and answer them any time during the test administration, offering examinees a sense of control and immunity to Wainer's strategy. Cui, Liu, He, and Chen (2018) suggest using the CAT with salt (CATS) approach to give examinees full control of item review and revision. This approach uses nonadaptive items in a test so that examinees cannot use Wainer's strategy to increase their scores.

Cui and colleagues (2018) suggest that the CATS approach has yet to be compared with other approaches such as the IP and MEOTR. Without knowing the performance differences of the three approaches, it is difficult for test practitioners to choose which approach to implement in a CAT administration. This study will compare CATS, IP, and MEOTR using simulation studies. All three approaches will have the same test length (30 items) and be based on the same item response theory (IRT) model (3PL). The comparisons also will be based on the same item bank containing 600 3PL items. Different conditions will be manipulated within each approach, as will the number of nonadaptive items (5, 10, and 15) for CATS; the pocket size, or number of items (i.e., 2, 3 and 5), for IP; and the number of revised items allowed (5 and 10) for MEOTR, respectively. This study will also employ regular, Wainer, and random test-taking strategies used in a previous simulation study (Cui et al, 2018). The bias, root mean square error (RMSE), and conditional standard error of measurement (CSEM) of the abilities will be compared among three approaches.

Measuring Nonverbal IQ in Autism: Challenges and Opportunities Building a Measure and Internet Administration Application

Author
John Hansen (Hansen Research Services LLC)
Jerel Unruh (Unruh Designs)

Abstract
Studies of Autism Spectrum Disorder (ASD) often include the measurement of intelligence for individuals on the spectrum as well as in genetically related siblings and parents. The wide range in ability among these participants as well as unique constraints such as limited attention span and young age make an adaptive measure appealing to researchers. In addition, because participants from this special population are most comfortable in their home environment, the Internet is an ideal tool for delivering the adaptive test administration.
Building the measure, a test administration system, and a data management system for researchers presented a number of challenges and opportunities for our team. This paper will discuss a number of salient questions encountered and decisions made during the engineering of the instrument, its deployment, and management. Further, we will address the many opportunities associated with these data captured from our measure of nonverbal IQ, bank of 700 items, and 24,000 participants from around the world. A demonstration of the instrument will be provided as well as functionality of the researcher data interface, system administration controls, and application programming interface (API). Particular areas pertaining to CAT will include: item banking, calibration, scaling, pretest, stopping criteria, item selection, step interval rules, content balancing, item exposure, and test security. Areas specific to ASD and children in general will include: interactions between item difficulty, entry item, response time, pacing, identifying participant fatigue, and test engagement. Lastly, because ASD is often comorbid with other types of intellectual disability as well as having differential symptom presentation by sex, questions of validity will be posited.

Developing and Piloting an Item Bank for a Computerized Adaptive Test of Adaptive Behavior Skills

Authors
Mo Chen, PhD (National Institute of Education, Nanyang Technological University, Singapore)
Kenneth Poon, PhD (National Institute of Education, Nanyang Technological University, Singapore)
Yong Hwee Nah, PhD (National Institute of Education, Nanyang Technological University, Singapore)
Huichao Xie, PhD (National Institute of Education, Nanyang Technological University, Singapore)
Vahid Aryadoust, PhD (National Institute of Education, Nanyang Technological University, Singapore)
Nicolette Waschl, PhD (National Institute of Education, Nanyang Technological University, Singapore)

Abstract
Adaptive behavior refers to the effectiveness and degrees to which a person meets the requirements of personal independence and social responsibilities, and it is one of the key criteria for the diagnosis of intellectual disability. The time required for the administration and completion of the common measures of adaptive behavior skills can range from 20 to 60 minutes, which could place participation burden on the respondents (e.g., parents, caregivers, teachers). In addition, there remains a general paucity of measures of adaptive behavior that are developed and normed in non-Western countries.
To address the above two potential issues, the current study as the first stage of a larger project is aimed to develop and pilot an item bank of adaptive behavior skills in Singapore context, which serves to the development of a computerized adaptive test (CAT) for the adaptive behavior measure. Specifically, five steps were taken to develop, pilot, and finalize the total 310 items, including 1) concept development, 2) initial item development, 3) item reviewing, consolidation, and refinement, 4) item distribution/assembly, and 5) piloting and item finalization. The finalized items have been currently used to collect data on a larger scale for item calibration and CAT development.

The Technical Underpinning of Physics Playground

Authors
Seyedahmad Rahimi (Florida State University)
Valerie Shute (Florida State University)
Russell Almond (Florida State University)

Abstract

Physics Playground (PP) is a 2D game with simple game mechanics (i.e., drawing objects or manipulating physics parameters to guide a green ball to a red balloon; see Figure 1). Through an iterative process, we designed various game levels and included multiple learning supports in PP to measure and improve students' physics understanding (Figure 2). We will test three different versions of PP across in May 2019: (1) linear order, (2) free-choice by students, and (3) adaptive selection of the next best level based on students' level of physics understanding. We are currently integrating an adaptive algorithm (Shute, Hansen, Almond, 2008) into PP which is based on the expected weight of evidence (Madigan & Almond, 1996) per task (i.e., level) as our level selection method. In this presentation, we describe the technologies used to implement both the player and administrative interfaces to the game and their communication with our assessment system. We will also share the important lessons learned.

An Urn-scheme to Track Accuracy and Response Times in Learning Environments

Author
Benjamin Deonovic

Abstract
An observed response can be conceptualized as an outcome of a match (with a certain duration) between an item and a player. In modelling these observed responses, it is usually assumed that person abilities are not changing -- at least not while making a test. If the skills of players or the difficulty of items possibly change over time (even after every response), this classical assumption of a constant ability fails. This is, for example, the case in learning environments with instructions or feedback.
This presentation proposes an algorithm based on urns that is designed to track changing parameters. The skill of each player and the difficulty of each item is represented with an urn with a certain probability (configuration of balls) and a certain size. This results in binomial distributed random variables with known standard errors. Second, we illustrate the urnings algorithm by analysing data from a large online adaptive learning environment (Klinkenberg, Straatemeier & Van der Maas, 2011) including both accuracy and response time data. We focus on the (developmental) relation between estimates based on speed and accuracy. Third, we show that this system allows for an adaptive selection of items to the skills of players, without any rating inflation.

History of CAT

Effect of Routing Errors on the Psychometric Properties of Multistage Tests

Authors
King Yiu Suen

David Weiss (University of Minnesota)

Abstract
As multistage testing (MST) is gaining popularity, it is important to know whether its measurement performance. The purpose of the current study is threefold. First, to examine how routing errors affect MST performance. A routing error occurs if an examinee is branched to a module that is not well matched with his/her true ϑ level. Second, to investigate how certain MST design factors affect the measurement performance of MSTs in the presence of routing errors. Third, to compare the measurement precision and efficiency of ϑ estimation produced by item-level CATs and MSTs designed for measuring individual differences, with and without routing errors in the MSTs.

Symposium: CAT in TURKEY : From Past to Present

CAT in Turkey : An Introduction
Alper Şahin (Middle East Technical University Northern Cyprus Campus)
This is a short introductory presentation about the history of CAT in Turkey, the challenges that the Turkish CAT researchers have, a summary of the CAT research topics in Turkey, the CAT centers that has been established in Turkey so far and a descriptive analysis of the CAT studies in Turkey. This introductory presentation will be followed by some fellow Turkish researchers in order to provide the audience with some sample CAT research from Turkey.

Development of the CAT version of a Turkish Interest Inventory
Eren Can Aybek (Pamukkale University)
R. Nükhet ÇIKRIKÇI (İstanbul Aydın University)
This study aims to adapt an interest inventory (Kuzgun, 2014) which is used widely in high schools in Turkey to a computerized adaptive test form. The interest inventory was meant to assess high school students' perceived abilities, interests and vocational values. The inventory has 23 factors. There are 10 items for each factor and a total of 230 items. Students are expected to respond each item as A (never, 1 point), B (occasionally, 2 points), C (often, 3 points), and D (always, 4 points) and they can score themselves and identify their own profile according to their factor scores.
A post-hoc simulation study was held and a live CAT application was developed for the study. Post-hoc simulations have been executed with data collected from 1144 high school students. Firestar and R software are used for post-hoc simulations by the following rules:

Stopping rule: Minimum three items and .30, .40 and .50 standard error
First item selection: Theta = .00
IRT Models: Generalized Partial Credit Model (GPCM) and Graded Response Model (GRM)
Item Selection: Maximum Fisher Information (MFI), Maximum Expected Information (MEI), Minimum Expected Posterior Variance (MEPV), Maximum Expected Posterior Weighted Information (MEPWI)
Theta estimation: Expected a Posteriori (EAP)

According to the simulation results GPCM was chosen as the IRT Model; MFI as the item selection method and .40 standard error was chosen as the termination criteria for the live CAT form. The simulation study showed that mean number of the items used is 113, while the paper & pencil (P&P) form has 230 items; and the correlation coefficients between simulation θ and the P&P scores for 23 factors was found to be between .90 to .96.

Live CAT form has been developed by the researcher on the Concerto platform with the help of catR package and administered to 25 high school students. The P&P and CAT form of the interest inventory profiles are compared and it has been seen that the profiles are similar. Correlation coefficients between P&P scores and θ levels from CAT have been found to be between .45 to .55 for three factors and .63 to .88 for 20 factors. Results of the simulation study and the graphical profile similarities between P&P and live CAT forms shows that the Inventory can be applied as a CAT.

References : Kuzgun, Y. (2014). *Self-evaluation inventory manual.* Ankara: Nobel Publishing.

Developing a MCAT to Assess Students' Language Competencies*

Murat Doğan Şahin (Anadolu University)
Selahattin Gelbal (Hacettepe University)

The aim of this study was to develop a multi-dimensional computerized adaptive test (MCAT) that measures grammar and vocabulary knowledge as well as general ability of the students attending school of foreign languages of a university in Turkey. To achieve this purpose, an item pool was created by using four different data sets each of which consisted of 50 questions used in the previous proficiency exams administered at school of foreign languages of this university. According to IRT-based analyses applied to the data, it was found that three-dimensional bi-factor model was the most appropriate model for all item sets. Following the development of the item pool, a hybrid simulation was applied –by taking the missing data into consideration – in order to determine the capability of the algorithm to be used in real-time MCAT application. In this process, a total of 36 conditions were created in which different ability estimation methods (EAP and MAP), different item selection methods (D-rule, KL, W-rule, T-rule, weighted W-rule and weighted T-rule) and different termination rules (standard error, θ convergence, fixed number of items) were tested. In these conditions, the correlation between estimated and true θ values for each dimension; bias, RSMD and standard error values were calculated. Being an important indicator of measurement accuracy in CAT applications in addition to correlation values and error statistic, the number of items exposed was also reported under the variable-length CAT simulation conditions. According to the results obtained, it was found that D-rule and MAP were the best methods for item selection method and ability estimation in each termination criteria. Following the comparisons of best conditions for each termination criteria, it was decided that the termination criteria based on standard error as well as D-rule item selection method and MAP ability prediction method should be used for real MCAT application. The minimum and maximum number of items that a test taker can take was set to 10 and 60 respectively. Then, the CAT developed was administered to 99 real test takers who were through the end of their preparatory school year. 32 out of 99 students took paper and pencil version of the test and the correlation between their P&P test scores and CAT scores were found to be .72. In addition, the total number of items administered to each individual and how frequently the items in item pool are used were reported. According to the results obtained from real-time application, it was found that 30% of the item pool, which included a total of 200 items were used. It was also observed that the average number of items replied was 12.3 and nearly 75% of the participants replied 10 to 12 items.

*** This study is based on the doctoral dissertation of the researcher at Hacettepe University Institute of Educational Sciences.**

Detection of cheating behavior in online unproctored CATs via a validation test

Ebru Balta (Ankara University)
Arzu Uçar (Ankara University)
Alper Şahin Şahin (Middle East Technical University Northern Cyprus Campus)

Computerized Adaptive Testing (CAT) is gaining much more attention than ever by institutions performing large-scale assessment, especially the ones attracting students worldwide, due to the increased security of CAT. The recent trend among some institutions, especially some universities, is that the overseas candidates are first assessed with unproctored CAT via internet. Then, the candidate applications are processed using the score of this unproctored test, and when the candidates arrive on campus they are asked to take a proctored non-CAT validation test which is parallel to the CAT taken online.

The challenge here is the detection of the cheating behavior based on the data of these two tests. The purpose of this study is to investigate potential solutions to this issue. For this purpose, first, a CAT dataset will be simulated for a number of test takers, then Paper & Pencil (P&P) test data will be simulated based on the ability estimates obtained from CAT simulations. Then, the data simulated for P&P will be modified by randomly selecting some candidates and they will underperform in P&P version of the test as some of their correct responses to some items will be replaced with incorrect ones. Their ability levels will be estimated one more time using the modified P&P data. Examinee ability levels in both tests will be compared via Kullback-Leibler divergence index. Then, in order to detect previously marked cheaters using the person-fit statistics (l_z) will be used and the type-1 error rate of person-fit(lz) index in identifying the cheaters will be calculated. The number of cheaters will be also altered in some conditions and the type 1 error will also be recalculated for different number of cheaters.

Data generation for P & P test will be performed using "*irtoys*" package of R. The data generation of the computerized adaptive test will be carried out using "catR" package. 100 replications will be performed for each condition in the generated data. Analyses for cheating detection will be carried out using the codes written by the researchers and the "*PerFit*" and "*CopyDetect*" packages in of R.

The Development of Theoretical Knowledge Test for Driving License Exam as a Computerized Adaptive Test*

Nükhet Çıkrıkçı[1], Seher Yalçın[2], İlker Kalender[3], Emrah Gül[4], Cansu Ayan[2], Gizem Uyumaz[5], Merve Şahin Kürşad[2], Ömer Kamış[2]

This study tested the applicability of the theoretical knowledge test for Candidates of Driving License (ECODL) in Turkey as a computerized adaptive test (CAT). The study was performed in three stages. Firstly, new item bank was piloted in real e-exams administered by Ministry of Education (MoNE). After pilot study, items were calibrated and selected with IRT based model. In the second phase, IRT calibrated item bank in various simulation conditions were tested for the live CAT. In this stage, results of the simulation studies were used to determine the rules for starting, continuing, and terminating the live CAT exam for ECODL. In the third stage of this study, the live CAT exam was applied according to the results of the simulation. Candidate drivers (n = 280) who had taken the ECODL as an e-test participated in the third stage. These 280 volunteers were asked to complete the live CAT. Moreover, they answered a questionnaire, which is surveyed opinions of candidate drivers on computerized adaptive testing. According to results of live CAT application, in terms of the rule of continuation of CAT application, expected a posteriori method gave the minimum measurement error. As a termination rule of CAT application, a fixed number of questions yielded the least

standard error of measurement. The estimated proficiency of the participants ranged from 0.35 to 0.56 in aspect of distribution of error (standard error of measurement). The calculated test information values varied from 3.16 to 8.21. It was also found that when ECODL was implemented as a CAT, it could reliably differentiate among testers in terms of competence of theoretical knowledge of driving and provide a basis for accurate decisions regarding their proficiency. Furthermore, according to the findings, although the ECODL_CAT had a fixed number of questions in this study, it was practical in terms of time taken to complete the test. Because ECODL_CAT considered the level of difficulty of questions in terms of individual proficiency, candidates did not have to deal with questions above or below their levels of proficiency. Moreover, according to questionnaire results candidate drivers do not have difficulty about CAT application technically and they think that CAT application is more practical and user-friendly than paper-pencil tests.

1. *İstanbul Aydın University, Turkey,* ***2.*** *Ankara University, Turkey,* ***3.*** *Bilkent University, Turkey,* ***4.*** *Hakkari University, Turkey,* ***5.*** *Giresun University, Turkey*

[a]This research (Project # 215k018) was supported by the TUBITAK (Turkish Scientific and Technical Research Institution) Project 1001 program in Turkey.

Symposium: Adaptive Learning and Adaptive Assessment: Intersections and Synergy

Authors
Michelle Barrett
Angie McAllister
Benjamin Deonovic
Krista Mattern
Drew Hampton

Abstract
Computerized adaptive learning and adaptive assessment systems have evolved substantially over the past 30 years. However, development and evaluation of these systems has often occurred independent of the other.
Before computing technologies were available to scale adaptive technologies widely, yet seeing impending potential, distinct academic communities began to work on adaptive technologies for learning and assessment. In accordance on the desired goal (e.g., increased learning efficiency, improved precision), and the underlying theoretical and related statistical models typically used by each community (e.g., cognitive sciences, learning sciences, educational psychology, psychometrics), extensive research to optimize adaptive methods ensued. Within each community, the definition of adaptivity became intimately intertwined with the optimization goal.
When computing power boomed, encouraged by digital disruption in other industries, a third community, Ed Tech, turned attention to the personalization of learning. A large number of education products became “adaptive”, albeit sometimes well outside the formal definitions of

adaptivity cultivated by the cadres of researchers focused on adaptive learning and assessment systems. As this happened, divisions between the adaptive learning and adaptive assessment communities seem to have deepened as each protected their own brand of adaptive.
Yet we likely have much to gain from deeper exploration of the intersection of these two worlds. Increasingly, collaborations among researchers in learning sciences, learning analytics, psychometrics, computer sciences, and data sciences are forming. Publications have emerged establishing underlying connections in the theoretical and statistical models used. Open-source software is increasingly available in both adaptive learning and adaptive assessment that exposes the "black box" and allows us to discover and build upon the work of one other.
We also face similar challenges. While learners and educators today minimally expect some degree of personalization, wide-spread adoption of scientifically-proven adaptive methods is elusive in a number of ways. Beyond optimizing our algorithms in small or simulation contexts, we are compelled to engage in the educational ecosystem at large. Industry standards and effective measures of efficacy at scale are next up, along with expanding the domains in which adaptive learning and assessment can be successful.

Symposium: Identifying Rushing in the *i-Ready Diagnostic* Computerized Adaptive Test

Authors
Laurie Davis
Can Shao
Bess Patton
Logan Rome

Abstract
Developing Item-Level Rush Flags for the i-Ready Diagnostic Assessment (Bess Patton, Logan Rome, Curriculum Associates)
The development of item-level rush flags was the first stage of a project to implement research-based rush criteria for the *i-Ready Diagnostic*. The *i-Ready Diagnostic* is an interim computer adaptive test taken three times a year by students in kindergarten through twelfth grade in reading and mathematics. To attain test scores that validly indicate what a student knows and can do, students must exhibit motivated and effortful behavior throughout the testing event. Therefore, it is pivotal to accurately identify non-effortful behavior.
One approach to conceptualizing non-effortful behavior is to identify rapid-guessing, e.g. rushing, at the item-level using response time-based measures. Time-based thresholds are identified for separating rushing from desirable effortful behavior on items. Prior research on the *i-Ready Diagnostic* indicated that a common threshold across items and grades may not be adequate for operational use. For example, a common threshold is too conservative for items that contain lengthy reading passages. Thus, item-specific thresholds were used to identify rushing.
The item-level thresholds were constructed using a modification of the *Normative Threshold (NT)* method (Wise & Ma, 2012). Current research advocates for the use of a 10 percent threshold of the mean response time for an item. However, prior to selecting a method for constructing the item-level thresholds, the stability of the mean and median was examined across testing windows. Research indicated that thresholds based on the median were more stable than those based on the mean, with 11% less items resulting in thresholds that would classify different students as rushing in both reading and math.

Thresholds were generated using 5-15% of the median response time (median NT05-NT15). These thresholds were evaluated by examining the distributions of the response times, percent of students flagged for rushing, and percent correct. First, a random sample of items was selected from each subject, domain, and grade level. Median NT05-NT15 thresholds were mapped onto the response time distributions for each item. The expected distribution contains a trough where the distribution of rushers intersects that of examinees exhibiting effortful behavior (Schnipke, 1995; Wise & Kingsbury, 2016). Visual inspection of 480 items indicated that median NT10-NT15 thresholds were reasonable, while NT05 was too low for most items.

Finally, students were flagged as rushing using the median NT05-NT15 thresholds. The percent of students flagged for rushing was evaluated in terms of face validity. Results confirmed commonly held beliefs, such as older students are more likely to rush than their younger peers. Due to the adaptive nature of the *i-Ready Diagnostic*, the expected percent correct was 50% for students exhibiting effortful behavior and chance (either 25% or 33%) for students rushing. An analysis of the percent correct revealed minute differences across thresholds. The desire to balance false positives and negatives coupled with the analyses results led to the selection of the median NT10 threshold.

Developing Test-Level Rush Flags for the i-Ready Diagnostic Assessment (Logan Rome, Bess Patton, Curriculum Associates)

The development of test-level rush flags for the *i-Ready Diagnostic* builds off the item-level flags described in the first talk. The test-level flags utilize Response Time Effort (RTE), the proportion of items on which the student was not flagged for rushing. Additionally, percent correct is used to check that students are indeed rushing, not answering quickly with effort. Two sets of rushing thresholds were created: a yellow flag used to alert educators to possible rushing; and a red flag used to identify severe rushing likely to impact the validity of the test score.

The test-level flags were decided by examining frequency distributions of RTE and percent correct, and test-retest correlations and extreme growth patterns after removing rushers using various criteria. After deciding on the flags, additional analyses were conducted for further validation. These analyses included comparisons of: student placement levels among rushers and non-rushers to detect bias of the rush flags, the standard error of measurement (SEM) for rushers and non-rushers, and the means and standard deviations of test scores prior to and after removing rushers from the sample. All analyses were conducted separately for two subjects (Reading and Math) and 13 grades (K-12); thus, 26 sets of analyses were used to decide on the flags.

Students in upper grades (middle and high school) tended to show lower RTEs than students in lower grades; this indicates decreased effort in upper grade students. Consequently, the average percent correct tended to decrease in upper grades; however, percent correct did not deviate far from 50%, as is expected in a CAT. Test-retest correlations consistently increased as more students were removed. That is, the most conservative flags yielded the best correlations.

Five yellow flags and six red flags were chosen for further inspection. From these, the final yellow and red flags were selected based on the growth patterns observed among rushers. Students who rushed on the initial test but not the retest showed growth that was three times that of their peers. On the other hand, students who rushed on the retest but not the initial test showed extreme negative growth due to the deflated scores on the second test. Based on these results, the following flags were chosen: Red – 0.75 RTE and 43%; Yellow – 0.90 RTE and 45% correct.

Further validation of the flags showed that the distributions of student grade-level placements were similar for rushers and non-rushers; the flags did not disproportionately identify students based on ability. Students who were flagged for rushing also had higher SEMs than non-rushers of similar ability. Finally, removing rushers from the sample increased mean test scores and decreased the standard deviation of scores; a sign that construct-irrelevant variance was removed.

This study, combined with the item-level study, provides a framework for testing organizations to investigate rushing in CATs. The results demonstrate how operational testing data can be used to set thresholds for flagging rushing. Finally, this study improves on prior research by considering percent correct, in addition to RTE, in rush flagging.

Comparison of Three Methods for Detecting Test-Level Rushing in the i-Ready Diagnostic Assessment (Can Shao, Logan Rome, Curriculum Associates)

This study proposes to use three methods to detect test-level rushing behavior in the *i-Ready Diagnostic* Assessment. These three methods are: 1. Response Time Effort (RTE) + percent correct method described in Talk #2; 2. Change-point analysis (CPA) with item response; 3. CPA with response time.

CPA is a likelihood-based method which serves as a well-established tool to check for abrupt changes in a sequence of data. It has been widely used in biology, statistics and economics. Because of its flexibility and robustness, CPA was first introduced to the field of aberrant response detection to flag examinees with speeded responses (Shao, Li, & Cheng, 2016; Sinharay, 2016). Unlike many person fit statistics which only work at the examinee level, this method can detect both the rushed examinees as well as the point at which they start to rush (see Figure 1 for two representative examinees). The key idea is to test the existence of the change point and pinpoint its location by plugging the estimated examinee abilities (before and after each possible change point) in the likelihood function. This method is further extended to use response time for speededness detection (Shao, 2017). With response time data, the examinee's working speed parameter in the log-normal model (van der Linden, Scrams, and Schnipke, 1999) can be estimated at each possible change point and checked for speededness using either Wald test or likelihood ratio test. So far, no study has been conducted to evaluate the performance of these two methods on the same dataset in terms of detection consistency of rushed students and where they start to rush. This will be demonstrated in the study as well as detection consistency with the RTE + percent correct method from Talk #2.

In this study, a sample of 194,523 students' raw response and the corresponding response time data for 66 items each were randomly drawn from the testing population. The CPA analysis using item response and response time were carried out in R. To obtain the critical value of the loglikelihood, the null distribution was constructed using permutation (i.e., randomly shuffling the sequence of response/response time for an examinee). Five thresholds (p-value=.05, .025, .01, .005, and .001 respectively) were picked based on the permutated null distribution to control for Type-I error rate. classification consistency statistics were calculated between the two CPA methods with different *p*-values and the RTE + percent correct method with different thresholds.

By evaluating different rushing behavior detection methods, this research aims to give practitioners some insights in choosing a proper method for related research. All three methods can be used either as a first step to check if rushing occurs during the test administration or to serve as a tool to flag rushing live (i.e., during the test) and remind students that they may need to pay more attention to the test. This is informative to teachers so that they can work with targeted students regarding the rushing issues.

This study details the process of implementing item-level rush flags operationally and highlights issues, considerations, and potential solutions for practitioners. Additionally, this study builds upon prior research by introducing thresholds derived from the median response time.

Getting Started with CAT: A Case Study

Authors
Jordan N. Stoeger
Tianjiao (Tina) Yin
Xiaoqiu Xu

Abstract
Computerized adaptive testing offers to make tests shorter, more efficient, and more accurate, but the mental and financial resources required to complete such a process are prohibitive for most organizations.
First, we will review the previously developed five-step model for developing a CAT (Thompson and Weiss, 1999). Then we will explain the tasks involved in a real-world application of the model. A Beijing startup leading the international market in virtual ESL teaching, VIPKid has been testing and placing students into their tiered curriculum since 2013. Utilizing this data, we aim to improve the existing placement exam by utilizing assessment best practices.
With the goal of making CAT more widely available this project utilizes reasonably priced and readily available software when possible. Where it is not, we offer suggestions for cost-saving methods. We hope to provide a groundwork for other organizations hoping to move toward CAT.

Is Your Item Pool Ready for the Demands of CAT?

Authors
Ryan A Wilke
Cody Diefenthaler

Abstract
This presentation involves a demonstration of a freely available web-based software tool, WebCATSim, that can be used to investigate the interplay between item supply and test demands within a simulated CAT environment. Test developers and researchers with a concern for CAT feasibility and the need to develop quality CAT examinations will have an interest in this presentation. We explore the tool's ability to evaluate such questions as: 1) Will my item pool work in a CAT environment? 2) What item selection algorithm might work best? 3) What percent of my items may be overexposed, given my test assembly rules? 4) Will my test have content balancing issues? Participants are encouraged to bring their laptops and explore item pools and CAT performance issues with us.

A primary goal of computerized adaptive testing (CAT) is to efficiently estimate a person's proficiency score (usually denoted θ or "theta"). Within a CAT, each subsequent item is chosen by an algorithm to precisely measure each examinee's provisional theta score (Lord, 1977), while also balancing the test demands with the supply of items in the pool to ensure CAT feasibility (Stocking, 1987; *Authors, 2017*). An inadequate inventory of items may result in several undesirable outcomes including the inability to meet the stopping criterion within particular regions of the scale or overexposure of those items with characteristics in high demand (Way, 1998).

Item pool development has been important to those interested in a CAT environment (Veldkamp & van der Linden, 2010). Test demands and item supply each play a unique role in determining CAT feasibility for different types of tests. For example, a testing program may wish to target measurement precision at a specific ability level or ensure stopping criteria are met efficiently. However, meeting such test demands is dependent upon the supply of items in the pool. Characteristics of the item pool including total number of items available, difficulty and discrimination of the items, number of items available within the vicinity of a specific proficiency level, and distribution of items across the entire proficiency range each contribute to the ability of the item pool to meet the test demands.

Predicting in advance of test administration the degree to which an item pool can meet test assembly specifications and precision requirements would seem to be extremely useful. As Thompson and Weiss (2011) noted, having good information about the limitations of an item pool and various trade-offs would be valuable to test developers, testing program administrators, and clients.

By modeling the CAT item selection process across the proficiency scale (that is, at designated θ values), WebCATSim evaluates the feasibility of the item pool to achieve the test's demands. Specific item selection algorithms, content balancing constraints, or item exposure limitations can be manipulated. The program executes CAT simulations employing item sequencing algorithms (Reshetar, 1990) across the proficiency scale at specific intervals of theta and returns both output files as well as graphical representations, which can be used for both visual inspection and reporting.

The MST Role To Reduce Public Cost In Testing

Authors
Mariana Curi (University of São Paulo, Brazil)
Thales Akira Matsumoto Ricarte (University of São Paulo, Brazil)
Alina Von Davier (ACTNext by ACT)

Abstract
In this work, we simulated a Multistage Testing (MST) application to the 2017 brazilian National High School Exam (ENEM, in portuguese) data. Our purpose was to illustrate the viability of reducing the number of days of the test administration based on a MST methodology, to save public resources, maintaining the quality and reliability of the test results.

ENEM is a non-mandatory national exam, held annualy, designed to evaluated high school students in BRAZIL in 1998. Since 2009, it has become one of the most important high-stake tests in BRAZIL, once it has been used as admission test for enrollment in most brazilian public universities. It is the second largest in the world (after the chinese National College Entrance Examination, NCEE), with around 7 million candidates per year,

composed of 180 multiple choice items, related to 4 competencies (1. language, 2. human sciences, 3. natural sciences, and 4. mathamatics) plus a written essay. It is applied in two days (like NCEE, in most provinces of China), the first lasting 330 minutes (competencies 1, 2, and written essay), and the second lasting 300 (competencies 3 and 4), in a paper-and-pencil format, with a total cost for the government almost equal to US$180 millions (cost announced for 2018 edition).
In our study, we adopted the same Item Response Theory (IRT) method used in the official ENEM, i.e., the unidimensional 3-parameter logistic model for each compentency. We also considered three stages schema for the MST (1-2-3), and the Fisher information function for routing the individuals through stages. The first stage module was composed with 10 items with a wide range of difficulty; the two modules of the second stage were also composed with 10 items each, and defined as easier and harder modules, respectively; and the third stage modules had 5 items each, specified as easy, moderate and difficult. With this scenario, our simulations were conducted fixing a 25-item adaptive test for each competency. A 5% sample over the total 6.8 million candidates of the 2017 edition was selected for item calibration and construction of the MST modules. For 50,000 random selected candidates of this sample, the estimated latent traits were calculated considering both the full length and our MST schema tests. The Spearman correlations between these estimations varied from 0.85 to 0.91 for the 4 competencies.
These results are promising to the propose of transforming large scale assesments to an adaptive design, shortening the number of application days. Reducing ENEM´s size from 180 to 100-item test enables the application in only one day, reducing costs and human resources considerably. In addition, the use of a MST scheme allows computerized or paper-and-pencil applications.

A Top-down Design Paradigm of Computerized Adaptive Multistage Test

Author:
Xiao Luo

Abstract:
Computerized adaptive multistage testing (MST) has been gaining increasing momentum recently. Different from the item-level computerized adaptive testing (CAT), MST administers pre-constructed modules over multiple stages and adapts the test at the stage level. The coarser adaptability slightly decreases testing efficiency in MST, yet it exerts greater controls on the content of the test than CAT. Concerning the test development, MST also adopts a different approach from CAT. The regulation of the psychometric and content requirements is largely embedded in the item selection rule in CAT so that the test forms would comply with those requirements when assembled on-the-fly. Conversely, these requirements are holistically considered in pre-administration test construction phase in MST. Test construction in MST, thus, becomes a very important aspect of MST.
A widely used test construction method is the bottom-up approach where the overall test requirements are divided across stages into module requirements. Modules are subsequently assembled individually to meet those requirements. This approach requires a sensible partition of the overall test requirements. An optimal partition is difficult to find in reality, especially for MSTs with a large quantity of constraints. This study introduces a top-down approach that allows the test to be designed holistically at the test level and uses an optimization algorithm to find the optimal partition.

The method will be elaborated in more technical details in the final presentation. However, at the higher level, this method includes the following steps. First, it maps each route with a target subpopulation. Second, test information function (TIF) is maximized over the target subpopulation for each route. Third, constraints are imposed on each route. Lastly, the test is assembled using a mixed integer programming (MIP) framework.

A simulation study will be conducted to compare top-down MSTs against bottom-up MSTs. The preliminary analysis found that routes targeting the same subpopulations were more homogenous in the top-down approach than the bottom-up approach and the routes targeting different subpopulations were more separated in the top-down approach. It means that the assembly outcomes were closer to the design intents in the top-down approach than the bottom-up approach.

After the assembly, the preliminary analysis also simulated the administrations of those MSTs to 10,000 test takers. It was found that top-down approach produced similar or lower root mean squared errors (RMSEs) between true and estimated abilities than the bottom-up approach. Furthermore, routes were used more evenly in the top-down approach. Together, the proposed top-down design approach shows a great potential of simplifying the test construction process for MST as well as improving the psychometric quality of MST.

Dual-Objective Adaptive Testing Using High-Order Item Response Theory Models

Author
Kuo-Feng Chang (University of Iowa)
Won-Chan Lee (University of Iowa)

Abstract
In traditional computerized adaptive testing (CAT), the main focus is on obtaining accurate estimates of individual proficiencies. Inspired by smart learning, modern educational assessments are transforming from measuring relative location of students' performance into more flexible and informative tools that can be used to address various other needs of stakeholders. One recent trend is to include classification results (e.g., basic, proficient, or advanced) of subdomain measures in report cards for diagnostic purposes. The current study proposes to use high-order item response theory (HIRT) models to achieve the above goal by implementing computerized classification testing (CCT) procedures in the first-order traits (i.e., subdomain proficiencies) to provide classification results as a formative assessment for diagnostic purposes while the estimates of higher order traits can provide overall performance for a summative assessment. It should be noted that under the current testing framework, the test is terminated once an examinee' subdomain abilities simultaneously satisfy the CCT termination criterion. The proposed procedure utilizes the advantages of HIRT models in obtaining both overall and subdomain proficiencies, and furthermore the strengths of CCT in gaining classification results for subdomain measures that can be used to support individualized learning.

In a simulation study based on the three-parameter logistic HIRT model (Huang, Wang, & Chen, 2010) with two-level latent traits (assuming the hierarchical structure consists of three first-order traits and one second-order trait) and multidimensional CAT (Segall, 1996) algorithms, the following factors are studied: (a) maximum test length—60 and 120; (b) factor loadings for the three first-order traits—high (.9, .8, and .7, respectively) and diverse (.9, .6, and .3, respectively); (c) item selection procedure—Fisher Information and the progressive method (Barrada,

Olea, Ponsoda, & Abad, 2008); and (d) CCT termination criterion—the sequential probability ratio test (SPRT, Eggen, 1999; Wald, 1947) only, ability confidence interval (ACI, Kingsbury & Weiss, 1983; Thompson, 2009) only, ACI plus the "standard error change" rule (ΔSE, Choi, Grady, & Dodd, 2010; Gershon, 2017) as a secondary stopping criterion (the joint stopping rule), and SPRT plus ΔSE. In addition, the number of cutting points is fixed at two for each first-order trait in this study. The Sympson-Hetter online method with freeze (Chen, Lei & Liao, 2008) and the modified multinomial model (Chen & Ankenmann, 2004) are implemented for the item exposure control and content balancing, respectively. Measurement efficiency for the second-order trait is assessed by the mean difference and root mean square difference of the trait estimates. Classification efficiency for the first-order traits is assessed by accurate classification rates, forced classification rates, and average test length. In addition, pool usage rate, item exposure rate, test overlap rate, and content balancing (defined as the percentage of selected items for each subdomain) are used to assess the effect of item selection constrains.

Multidimensional CAT Measuring Patient-Reported Outcomes in a Hospitalized Population

Authors
Yikai Lu
David J. Weiss (University of Minnesota)
Chun Wang (University of Washington)

Abstract
Wang, Weiss, and Shang (2018) developed, and studied in simulation, a compound stopping rule to reduce the length of a multidimensional CAT (MD-CAT) without negatively impacting its observed measurement error. The rule used a standard error termination criterion (from the Fisher information matrix) followed by a ϑ convergence criterion when necessary. The MD-CAT was then applied in live testing at the Mayo Clinic to assess self-report of cognitive functioning, daily activity, and mobility of 776 hospitalized patients. Results from the live testing were compared with those from the simulations. The major finding was that the vast majority of patients' MD-CATs were terminated by the standard error criterion, in contrast to the simulation results. In addition, results of the live CAT were compared between males and females on a number of variables, including ϑ estimates, number of items administered, observed SEMs, and testing times. More research on calibration of the compound stopping rule for use in live testing is suggested.

Implementing a Large-Scale Computer Adaptive Assessment System: The Role of Culture and History Learning

Author:
Stanley Rabinowitz

Abstract:
In addition to applying research-based, accessible solutions and strong management tools and oversight, any assessment system must reflect local culture and history as it is conceptualized, built and implemented. The presenter has managed the development of successful large scale national assessments in both the US and Australia assessing millions of students and thousands of schools annually. The session will focus on the following key decision points and how they played out in these different environments.

- Purpose: what is the assessment system 'theory of action?'
- Governance: how are key decisions made and carried out?
- Stakeholder involvement: timing and methods to ensure buyin;
- Population: full cohort or sample assessment?
- Stakes: for students, schools, systems (real and perceived);
- Implementation approach:
    - Phase in vs big bang;
    - Number of years;
    - Activities each year (eg, item and platform testing);
    - Timing of platform and content enhancements;
- Ensuring school/student/platform online readiness;
- Content: what subjects/domains should be assessed? How often (annual vs cyclical)?
- Item Types, including mix of traditional and technology enhanced items, use of video;
- Adaptive model: Item vs stage adaptive;
- Fairness: ensuring validity of results for SWDs, ELs and other at-risk populations, including accommodations, alternative items and bias review;
- Comparability: online vs paper, within and across years (impact on trend data);
- Use of auto-scoring (AI): how to frame, validate and build support;
- Reports/dissemination of results: how to balance accessibility with technical accuracy

Successful implementation will only occur if decisions made around these challenging issues truly respect and reflect local culture and history.

Test-takers' Psychological Aspects in Case of a Small-scale Multistage Computer Adaptive Testing

Author
Tetsuo Kimura (Niigata Seiryo University, Japan)

Abstract

Kimura & Nagaoka (2011) suggested that experience of taking CAT discourages students and may cause backwash effects such as loss of learning self-efficacy and motivation because most of them felt the test had been ”difficult,” and many felt ”discouraged” or ”unsatisfied” with the experience. Kimura & Koyama (2016) revealed that test takers tend to prefer the longer test with easier items than the shorter test with more difficult items based on questionnaire immediately after different CATs administration. Kimura (2017) implemented a multistage CAT system on Moodle and administered a two-stage English vocabulary CAT to Japanese university students. At the end of the first stage, they can choose “a shorter test with more difficult items” or “a longer test with easier items”. Most of them chose the former for their second stage. The purpose of the current study was to scrutinized psychological aspects of test takers more clearly by interviewing test takers after the CAT. They were interviewed as they reviewed their test taking process by watching video that recorded screen image during the test. Similar tendencies were confirmed as the previous studies and their comments suggested that “a longer test with easier questions” may have had a positive effect on their motivation and self-efficacy. They also had a positive impression of the testing system that they can choose the module at the end of the first stage.

The Development and Efficiency of the Computerized Adaptive Testing Version for the Test of Chinese as a Foreign Language. (TOCFL-CAT)

Authors
Po-Hsi Chen (National Taiwan Normal University, Department of Educational Psychology National Taiwan Normal University, Institute for Research Excellent in Learning Sciences)
Chia-Ling Hsu (The Education University of Hong Kong)

Abstract
The Test of Chinese as a Foreign Language (TOCFL) is a standardized Chinese certificate exam for the non-native speaker Chinese learners. TOCFL is developed by Steering Committee of the Test of Proficiency-Huayu (SC-TOP) since 2006 and proved by Taiwan government. It can be used for applying scholarships in Taiwan and as a Chinese language certificate when apply to college. TOCFL is carried out in more than 30 countries and 60 cities each year. There are about forty thousand test takers every year in the world. The computer based version of TOCFL divided into three bands of tests and can evaluate six grade of proficiency that align to Common European Framework of Reference (CEFR). Since many test takers felt it difficult to choice the suitable test bands for them to take, the SC-TOP began to developed computerized adaptive testing version of TOCFL (TOCFL-CAT) for listen and reading test. To evaluate the efficiency of TOCFL-CAT compared with non-adaptive TOCFL, a simulation study was carried out. Different items selection rules and stopping rules were used to be the independent variables. The number of items they took to reach the preset test precision level, the root mean square of error (RMSE) of the ability estimation and the percentage of correctly classify (PCC) were calculated to be the dependent variables. Results demonstrated that TOCFL-CAT can save more than 30% of items to reach the same precision level as the non-adaptive TOCFL. The maximum information item selection methods with content and item type limitation were

suggested to be used in TOCFL-CAT. The computerized adaptive testing version of TOCFL (TOCFL-CAT) listen and reading test had been carried out since 2016 in Taiwan and some countries in Asia, Europe and America.

Conversational Assessments for AI based Adaptive Testing

Authors
Dee Kanejiya (Founder & CEO, Cognii)

Abstract
As the AI technology transforms a number of industries, how can educational assessments benefit from this innovation? This session will provide an introduction to the conversational assessments that combines the pedagogy of conversation with the technology of conversation - natural language processing - to create a scalable high-quality education experience.
Conversation is at the core of education and training. In the school, the primary mode of instruction delivery and learning assessment is the conversation between the teacher and the students. In the workplace, knowledge is shared via a sequence of short conversations between employees. A conversational assessment involves open-response questions and instant formative feedback. Benjamin Bloom famously demonstrated that the students who receive instant feedback in a one-to-one environment perform two standard deviations better than students learning in a passive cohort. Similarly, many educational researchers and policy makers believe open-response (aka short essay) questions are the best types of assessment items for measuring students' ability to actively recall knowledge, think critically, and solve problems - skills that are important for succeeding in the 21st century work environment.
Despite the multi-faceted benefits, open-response questions remain very infrequently utilized in testing due to the time and cost constraints in grading the answers. Similarly, despite the relevance of brief conversations as a modus of knowledge exchange in real-world, most of the educational tests today contain multiple-choice questions requiring very little construction by students as they select a response from a given set of choices, and essay questions requiring a very lengthy construction by students. This bi-polar state of educational assessments is attributable to its origin during the industrialization era where the focus was on producing uniformly skilled workers for factories, which is very different from today's skill-diverse innovative work environments emphasizing creativity and problem solving. Due to this stark contrast between the nature of assessments and real-world skill demonstration, educational assessments are often perceived as artificial, stressful, and even invalid by people who experience a mismatch between their earlier academic performance and later life achievements.

Virtual Learning Assistant(VLA) is a new genre of AI based EdTech that supports the most efficacious conversational assessments at scale. Compared to general purpose virtual assistants such as Siri or Alexa that only focus on answering a user's question, or 'essay graders' that focus on generating a number given an answer via brute-force modeling, the VLA provides a unique solution of grading students' open-response answers with a qualitative feedback and guiding them towards conceptual mastery in an adaptive sequence. Conversational assessments enable high quality testing that is reflective of real-world scenario, without creating the stressful environment associated with typical testing.
This session will share insights from efficacy studies conducted in a number of schools and colleges using the conversational assessment technology, and will provide guidance on the future of educational assessments.

A machine learning "Rosetta Stone" for psychologists and psychometricians

Authors
Alan Mead (Talent Algorithms Inc.)
Jilin Huang

Abstract
Do you know how and why to perform "one-hot" encoding for a "co-occurrence matrix?" Or how and why to perform "2-fold" crossvalidation? What are "bagging," "boosting," or "features?" You will understand these terms after you hear this talk about a "Rosetta Stone" for psychometricians and psychologists to understand the jargon that data scientists are inventing, in many cases for concepts we have been using for decades. Amuse your friends by talking like a data scientist!

Feasibility of Using Small Samples to Train an Automated Essay Scoring Engine

Authors
Nathan Thompson (Assessment Systems Corporation)
Alper Şahin (Middle East Technical University)

Abstract
The automated essay scoring has a history of more than a decade. There are many examples of such engines finetuned by the large assessment companies. Those engines have reached to such a state that they can estimate the test taker ability levels much more accurately than human raters. This is mostly because of the learning ability of the AI engine, the unchanging criteria throughout the whole essay scoring process and obviously tireless machine scorers surpassing the intra-rater and interrater reliability of human scorers.

Developing such an AI engine to score essays automatically is a burdensome process as it requires careful coding of algorithms and extensive research. There are some steps to take while developing and finetuning such an engine. One of them is studies to identify the robustness of the scoring engine to the dramatic decrease in the number of previously scored essays used to feed and make it learn the scoring scheme in each individual assessment. This study is designed to measure the capability of the AI engine when samples of limited number of essays used to feed it. Thus, the aim of this study is to find out how feasible it is to use small samples to feed an AI engine. For this purpose, an automated essay scoring algorithm is coded in R taking Text2vec R package as a base. Then, essay data collected from N=1991 high school students is used to feed the engine. Then random samples of 1000, 750, 500, 350, 250, 150 are drawn from the data and their scores are estimated to see the correlation of these scores with their pseudo-true scores by the human raters. RMSD values are also calculated.
Then among the full data of essays, samples of 1000, 750, 500, 350, 250, 150 are drawn keeping the score distribution in the full data intact by utilizing stratified sampling method on SPSS. The samples of essays drawn are used to feed the AI engine individually and scores of the 1000, 750, 500, 350, 250, 150 test takers randomly selected at the beginning are estimated again after each time the engine is fed with stratified samples of essays with varying sizes drawn from the full data. The correlation and RMSD values between the essay score estimates of the same randomly selected sample of essays obtained in the first step where the full data is used to feed the engine and in the following steps in which the stratified samples with varying number of essays are used to feed the engine are calculated.

Psychometric evaluation of computerized adaptive tests to assess quality of life in people with diabetic retinopathy

Authors:
Eva K. Fenwick (PhD) (Singapore Eye Research Institute, Singapore National Eye Centre, Singapore; Duke–NUS Medical School, Singapore)
John Barnard (Excel Psychological & Educational Consultancy, Melbourne, Australia)
Aiden Loe (PhD) (The University of Cambridge Psychometrics Centre, The University of Cambridge, United Kingdom)
Alfred Gan (MSc) (Singapore Eye Research Institute, Singapore National Eye Centre, Singapore)
Jyoti Khadka (PhD) (Discipline of Optometry and Vision Science, Flinders University of South Australia, Australia)
Konrad Pesudovs (PhD) (Consultant, Adelaide, Australia)
Shu Yen Lee (Singapore Eye Research Institute, Singapore National Eye Centre, Singapore)
Gavin Tan (Singapore Eye Research Institute, Singapore National Eye Centre, Singapore)
Tien Y. Wong (PhD) (Singapore Eye Research Institute, Singapore National Eye Centre, Singapore; Duke–NUS Medical School, Singapore)
Ecosse L. Lamoureux (PhD) ((Singapore Eye Research Institute, Singapore National Eye Centre, Singapore; Duke–NUS Medical School, Singapore)

Purpose: To psychometrically evaluate “RetCAT”, ten diabetic retinopathy (DR) quality of life (QoL) computerized adaptive tests (CATs), in a multi-ethnic clinical sample of patients with DR.

Methods: Rasch-calibrated item banks were developed to measure DR-related SY-Symptoms; AL-Activity Limitation; MB-Mobility; EM-Emotional; HC-Health Concerns; SC-Social; CV-Convenience; EC-Economic; DV-Driving; LT-Lighting. CATs for each domain were developed by Excel Psychological & Educational Consultancy. Monte-Carlo simulations generated abilities for cohorts of 1000 hypothetical test-takers. For each domain, an initial simulation was based on normal (N ~ (0,1)) distributions with abilities in the interval [-3;3] logits. Precision (standard error (SE) of each ability estimate) was stepwise reduced as SE≤0.50; SE≤0.40; SE≤0.35; SE≤0.30 and SE≤0.25. Positively and negatively skewed ability estimate distributions were explored. As simulations suggested that most domains achieved SE≤0.35 with seven items, RetCAT was set to administer seven questions per domain (total=70). In a cross-sectional, clinical study of English and/or Mandarin-speaking Chinese, Malay or Indian DR patients recruited from retinal clinics at the Singapore National Eye Centre, participants were interviewer-administered individually timed RetCAT tests online (131 English; 52 Mandarin). We evaluated: 1) the efficiency of RetCAT (i.e. SEM achieved; time needed to complete each CAT); and 2) its psychometric properties, including a) test information function (TIF) to assess test precision (higher TIF=higher reliability) and identify where the tests had highest/lowest measurement precision; and b) item exposure rate (IER) to assess the proportion of items administered overall and ≥50% of the time. As EM scores were independently lower in Mandarin compared to English, we report the psychometric results for EM separately by language.
Results: Of the 183 participants (mean age±SD 56.4±11.9 years; 61% male; 66% Chinese), 132 (76%) and 41 (24%) had non-proliferative and proliferative DR, respectively. The mean SEM across all RetCAT domains was 0.351 (0.88 reliability), with SEM±standard deviation values ranging from 0.272±0.130 (0.93 reliability) for EC to 0.484±0.130 (0.77 reliability) for EM-Mandarin. The median time taken across all RetCAT tests was 1 min 47 secs, ranging from 1.12 (IQR [Interquartile range] 1.63) for DV to 3.28 (IQR 2.52) for AL (median total time taken=17 mins 17 secs). The TIFs for RetCAT varied, ranging from 6 (0.80 reliability) for small item banks (e.g. LT, n=10 items) to 48 (0.98 reliability) for large item banks (e.g. AL, n=92 items). Measurement precision was highest for participants at the lower end of the 'ability' spectrum (i.e. most impaired). The IER varied across the RetCAT tests; for example, VS, HC and EC had 100% IER, while CV and DV had <50% of available items administered. For most tests, 30-40% of items were administered >50% of the time.
Conclusions: RetCAT can provide efficient, psychometrically robust and comprehensive measurement of DR-related QoL impact. RetCAT may be useful to monitor patient progress, evaluate the patient-centred impact of new treatments, and undertake value-based evaluations of patient care in tertiary eye centres. Future work is needed to better understand the cultural differences between Mandarin and English versions, and determine the responsiveness of RetCAT to interventions associated with DR and vision impairment.

Evaluation of a Computerized Adaptive Testing System for the Impact of Vision Impairment Questionnaire (IVI-CAT)

Authors
Ryan EK Man, PhD (Singapore Eye Research Institute, Singapore National Eye Centre, Singapore; Duke-NUS Medical School, Singapore)
Bao Sheng Loe,PhD (School of Psychology, University of Cambridge, United Kingdom)
Jyoti Khadka, PhD (University of South Australia, Australia; Registry of Older South Australians, South Australian Health and Medical Research Institute; College of Nursing and Health Sciences, Flinders University)
Gwyn Rees, PhD (Centre for Eye Research Australia, Royal Victorian Eye and Ear Hospital, University of Melbourne, Australia)

Eva K. Fenwick, PhD (Singapore Eye Research Institute, Singapore National Eye Centre, Singapore; Duke–NUS Medical School, Singapore)
Ecosse L Lamoureux, PhD, (Singapore Eye Research Institute, Singapore National Eye Centre, Singapore; Duke–NUS Medical School, Singapore; Department of Ophthalmology, National University of Singapore, Singapore

Background: The 28-item Impact of Visiompairment (IVI-28) questionnaire is commonly used to assess vision-specific quality of life (VRQoL). However, it is in a paper-pencil format, time-inefficient, and requires considerable resources for data analysis and interpretation. In this study, we compared the efficiency of simulated computerized adaptive tests (CATs) versus a paper-pencil IVI-28-item questionnaire and its associated 20-item Vision-specific functioning (VSF) and 8-item Emotional well-being (EWB) subscales.
Methods: Using the paper-based IVI data pooled from 832 Australian participants with vision impairment (VI) across 8 studies, the item calibrations of the main IVI, VSF, and EWB were generated with Rasch analysis using Winsteps software and the Andrich rating scale model. With these calibrations, CAT simulations were subsequently conducted on 1000 cases using the Firestar-D software, with 'moderate' and 'high' precision stopping rules (standard error of measurement [SEM] of 0.521 and 0.387, respectively). We examined the average number of items needed to satisfy the stopping rules; as well as correlations between the paper-pencil IVI and CAT simulated person measures (higher score indicating greater correlations), Item Exposure Rates (IER; higher proportions translating to greater item use), and test precision using test information function (TIF; higher values indicating better precision) for both precision estimates.
Results: Of the pooled participants (mean age±standard deviation (SD), 75.8±13.2 years; 63.8% male), 67 (8.1%), 224 (26.9%), 393 (47.2%) and 147 (17.7%) had no (Logarithm of the Minimum Angle of Resolution (LogMAR) ≤0.3), mild (0.3 > LogMAR ≤ 0.48), moderate (0.48 > LogMAR ≤ 1.00) and severe VI (LogMAR >1.00), respectively. For the main IVI-CAT, 5 and 10 items were required, on average, to obtain moderate and high precision estimates of VRQoL, respectively, corresponding to an 82% and 68% reduction in items compared to the paper-pencil IVI. Similar results were obtained for the VSF-CAT (5 and 11 items, corresponding to item reductions of 75% and 50% for moderate and high precision estimates, respectively), although item reduction was observed only for moderate precision estimates for the EWB-CAT (6 items, corresponding to a 25% item reduction). Correlations were high between paper-pencil and CAT person measures, with correlation coefficients of 0.89 (moderate precision) and 0.96 (high precision) for IVI-CAT; 0.90 (moderate precision) and 0.97 (high precision) for VSF-CAT; and 0.97 (moderate precision) and 1.0 (high precision) for EWB-CAT. The IER for IVI-CAT (high precision), and VSF-CAT and EWB-CAT (moderate and high precision) were reasonably distributed, whilst it was skewed for the IVI-CAT at moderate precision, with only six of the EWB items used >40% of the time. Test precision ranged from moderate to excellent for all three CATs, (TIF [reliability]: 16.5 [0.94] for IVI-CAT, 11.0 [0.91] for VSF-CAT and 4.2 [0.76] for EWB-CAT).
Conclusion: Compared to the paper-pencil IVI, the IVI-CATs required fewer items, with high measurement precision, making them potentially attractive outcome instruments for clinical trials, healthcare, and research. Item administration skewing towards EWB items for the IVI-CAT were resolved with an item exposure limit rule added to the CAT algorithm. Final versions of the IVI-CATs are available electronically.

Assessing item-level fit within the multistage testing environment

Authors

Xue Zhang
Chang Wang

Abstract
Item-level fit analysis plays an important role in scale development, as it not only serves as a complementary check to global fit analysis, but also guides item revision/deletion (Liu & Maydeu-Olivares, 2014). For the purpose of item bank maintenance and management, assessing item-level fit is essential to govern retiring overexposed or obsolete items over time and replacing them with new ones. Multistage testing (MST) is gaining popularity for many large-scale computer-based testing programs, in which preconstructed sets of items are administered adaptively and scored as a unit. MST increases testing efficiency and improves measurement precision compared to traditional linear tests or linear parallel forms tests. Moreover, compared to item level computerized adaptive testing (CAT), MST can provide better test security, allow greater control over test construction, keep away from sparse operational data matrix (Stark & Chernyshenko, 2006). With multistage testing, adaptation occurs at the item set level. Hence, the response data were missing at random. Chi-square-based item fit indices (e.g., Yen's Q_1 , McKinley and Mill's G^2 , Orlando and Thissen's $S\text{-}X^2$ and $S\text{-}G^2$) are the most widely used statistics to assess item-level fit. However, they may not handle incomplete data easily in their current form, because of relying on total score for grouping whereas the total scores are no longer comparable across test takers. To this end, we propose modified versions of $S\text{-}X^2$ and $S\text{-}G^2$ to evaluate item-level fit within the multistage testing environment. Their performances are evaluated via simulation studies, the manipulated factors include item bank size, module length, number of stage, sample size, and misfit source. Throughout this study, the 'mstR' package (Magis, Yan, & von Davier, 2018) is used to construct MST tests. Maximum likelihood (ML) estimate and maximum Fisher information (MFI) are chosen as the ability estimation method and module selection method as default in 'mstR', respectively.

Using Multidimensional Item Response Theory for the Multistage End of Primary School Test

Author
Maaike van Groen

Abstract
The End of Primary School Test (College voor Toetsen & Examens, 2018) in the Netherlands provides an advice to teachers and pupils about the most suitable level of secondary education for the pupil. The Dutch secondary education system consists of five levels: basic, lower, and middle vocational education, general secondary education, and pre-academic education. The individual decision for the level of secondary education is based on a combination of the teacher's judgment and an end-of-primary-school placement test. This placement test encompasses the measurement of reading, language skills, mathematics and writing.

The Dutch parliament has decided to change the test format to multistage testing (Yan, Von Davier & Lewis, 2014). A major advantage of multistage testing is that the tailoring of the tests is related to the ability of the students. Also, classification accuracy is higher for multistage tests than for linear tests. A domain-specific multistage test is constructed for reading, language skills, and mathematics. All students will make the same first stage module. After this initial module, a routing decision is made. One of the following modules is selected based on the student's previous responses. This routing process is repeated multiple times.

The End of Primary School Test is currently modelled using a unidimensional item response theory scale per domain. This implies that each domain-specific multistage test is based on one unidimensional scale. These unidimensional scales are used to select the items, to make the routing decisions, to monitor test difficulty, to equate test forms, and to report.

An alternative approach has been suggested for modelling the End of Primary School Test. Multidimensional item response theory can be used when measuring multiple subjects. In a between-items multidimensional model, all items are scaled in one multidimensional model with each item loading on one dimension. This approach could possibly be an interesting alternative to the current multi-unidimensional approach. However, all aspects of the test development process need to be reconsidered when changing the test methodology.

The test provides an advice to pupil's about the most appropriate level of secondary education. The current methodology uses a weighting procedure to combine the multiple unidimensional scales into one advice. The weights in this procedure are determined using a prediction model based on the current educational level of pupils that took the test three years ago. A similar procedure could be used when using multidimensional item response theory.

The test is currently assembled by test experts using automated test assembly. Specialized software was developed by Angela Verschoor for selecting the most optimal items given content and psychometric constraints. This software needs to be adapted to select items within a multidimensional framework. Fortunately, previous research has shown that items can be selected per dimension when using a between-items multidimensional model.

The current study addresses some of the adaptations that need to be made in order to be able to use a multidimensional model. Further studies are needed before a decision can be made about the modelling framework of the End of Primary School Test.

Using Item Response Theory to Improve Educational Assessment in Nigeria

Authors
Oladele Babatunde Kasim
Adegoke Benson Adesin

Abstract
One of the core and compulsory subjects at all levels of education especially at the primary and secondary schools in Nigeria is Mathematics. Consequently, this subject is made compulsory to be passed at all levels of education where it is being offered. This study will assess the ability of students in constructed-response mathematics tests of West African Examinations Council (WAEC) and National Examinations Council (NECO) using test theories in order to ascertain the level of students' achievement in the subject. Survey design of descriptive research type will be

adopted for the study. The population will consist of senior secondary schools students in Ibadan metropolis of Oyo State in Nigeria. Simple random sampling will be used to sample 12 schools in Ibadan Metropolis and 770 students will make up the sample size. Data collected will be analysed using item analysis of IRT PRO 3. Results of the finding will provide examiners and stakeholders in educational assessment with the knowledge of how important the needs to establish psychometric properties of constructed-response items, analyse, interpret parameters of test items and score test items within a good test theory framework.

Adaptive Essay Test Assessor For Secondary School Economics

Authors
E.T 'Duke' Sowunmi (Department of Social Sciences Education, Faculty of Education, University of Ilorin)
H.O. Owolabi (Department of Adult and Primary Education Faculty of Education, University of Ilorin)

Abstract
Adaptive testing has been popularly used for objective tests. Though psychometric properties of multiple-choice tests are easier to obtain compared to that of essay test, it has become imperative for test developers to come up with adaptive tests using essay type items. The Essay Test Assessor has been developed and validated for use in the assessment of short response essay tests at the secondary school level. This study is designed to validate an
adaptive component based on the cognitive demands of each item on examinees with very few items on the test. The Adaptive Essay Test Assessor (AETA) will be validated in Economics among secondary school students in +Nigeria. An item pool of 70 short response essay test items, which is very few when compared to the traditional adaptive test items pool, will be developed and administered on a sample of second year senior secondary school
students in Nigeria. Relative standing of students exposed to the adaptive essay test and their scores on a parallel multiple-choice test will be analyzed to determine the extent to which AETA is validly gauging the ability of learners.

Scripted CATs via an Adaptive Blueprints with Multistage Considerations: A simulation study

Author
Ye Ma (The University of Iowa)
Johnny Denbleyker (Houghton Mifflin Harcourt)
Shuqin Tao (American Institute for Research)

Abstract

Adaptive tests have a feature that allows measuring examinees' ability more efficiently by sequentially selecting the item, or a set of items, which helps maximize the information in the ability estimator. However, various constraints such as content type, depth of knowledge (DOK), item format, etc. typically are incorporated with test blueprints, which can constrain the level of adaptation. There are several existing methods for constrained Computerized Adaptive Tests (CAT) including Item-pool partitioning, Weighted-Deviation Method, Maximum priority Index method, testlet-based CAT, Multistage Testing, and Shadow tests approach (Kingsbury & Zara,1991; Swanson & Stacking, 1993; Cheng & Chang, 2009; Wainer & Kiely, 1987; van der Linden & Adema, 1998; van der Linden, 2005).

This study proposes a different approach of a constrained CAT: an adaptive blueprint approach that attempt to takes advantages from features of existing methods. Specifically, an item selection table will be created considering several constraints including item grade, strand, DOK, Item format (e.g., TEI vs. MC) and item response time, which could be helpful control speediness effect in CAT simultaneously. The item selection table will also be divided into several stages and grouped into three difficulty levels in terms of item difficulty. Therefore, this item selection table incorporated with the item selection algorithm, such as maximum item information, will be used to form the constrained CAT.

Research goals

This study aims to evaluate the performance of the adaptive blueprint incorporating content attributes such as item grade, item strand, DOK, item response time by comparing it's performance with tradition approach, which specifically means meeting only content balancing in each domain based on the target Rasch difficulty without controlling for other content attributes. Numbers of constraints that the given item pool could afford is another research question investigated.

Method

Eight hundred items from an operational Grade 5 to 7 item bank for a Grade 6 assessment serve as the item pool (see Table 1). This test is fixed-length with 30 operational items. Important study factors include the following: (1) test assembly design (adaptive blueprint approach vs. traditional approach only using target Rasch difficulty with content balancing; (3) degree of constraints (2 constrains: item grade and strand (2C) vs. 4 constrains: 2C+DOK+TEI (4C) vs. 2C + response time (RT) vs. 4C + RT) To evaluate the assembled tests, 3000 simulations with true theta generated from a standard normal distribution will be generated for the assembled CAT tests. The maximum information rule will be used as the item selection algorithm. The root means square error (RMSE) between true theta and estimated thetas (Rasch model). Difference from two test assembly design with respect to the constraints will also be reported.

Significance and Implications

The primary purpose of this design is to develop some balance with more content attributes as an adaptive feature, which has an intention to increase the validity and transparency of reported test scores of reported test scores. This proposed methodology borrows from the advantages of module approach inherent with multistage testing. It guides the constraints in a more direct way that allows test development staff to review and consider various context effects as part of the test assembly process.

Advance in CD-CAT: The General Nonparametric Item Selection Method

Author

Chia-Yi Chiu (Associate Professor at Rutgers University)

Yuan-Pei Chang (Rutgers, The State University of New Jersey)

Abstract
Computerized adaptive testing (CAT) is characterized by its high estimation efficiency and accuracy, in contrast to the traditional paper-and-pencil format. CAT specifically for cognitive diagnosis (CD-CAT) carries the same advantages and has been seen as a tool to advance the use of cognitive diagnosis assessment for educational practice. A powerful item selection method is the key to the success of a CD-CAT program and to date, various parametric item selection methods have been proposed and well-researched. However, these parametric methods all require large samples to secure high-precision calibration of the items in the item bank so that for each examinee, the selection of the most appropriate item is guaranteed. Hence, at present, implementation of the parametric methods to small-scale educational settings like the classroom remains challenging. In response to this issue, Chang, Chiu, and Tsai (2018) proposed the nonparametric item selection (NPS) method based on the nonparametric classification (NPC) method (Chiu & Douglas, 2013).
The NPS method does not require any parameter calibration and outperforms the parametric
methods for settings with only small or no calibration sample. Nevertheless, the NPS method is not without limitation. Wang and Douglas (2015) proved that the estimator of examminees' attribute profiles obtained by using the NPC method is statistically consistent if the probability of a correct response for examinees who master a subset of required skills is less than 0.5 and that for those who master all required skills is greater than 0.5. These assumptions are easy to fulfill when data conform to a conjunctive cognitive diagnostic model (CDM), such as the deterministic input noisy output "AND" gate (DINA) model (Junker & Sijtsma, 2001; Macready & Dayton, 1977), but may become stringent when data conform to more complex CDMs. Not surprisingly, the same restriction is carried over into the NPS method, resulting in unsatisfactory performance when the assumptions are violated.
To remedy for the shortcoming, the general nonparametric item selection (GNPS) method that incorporates the newly developed general NPC (GNPC) method (Chiu, Yan, & Bian, 2018) as the classification vehicle is proposed in the study. The GNPC method has a built-in mechanism that automatically adjusts the implemented weights according to the CDM underlying the data and thus, relaxes the assumptions imposed to the NPC method. The flexibility of the GNPC method allows for the development of the GNPS method that consequently can be used with any CDM or multiple CDMs without abandoning the advantage of being a small-sample technique.
The performance of the GNPS method with small calibration samples was evaluated by using
simulations. The results showed that the GNPS method outperformed several frequently used
parametric methods, providing evidence that the GNPS method is a more appropriate tool of cognitive diagnosis for small educational settings.

A Nonparametric Approach to Computerized Adaptive Testing

Author
Soleyman Zolfagharnasab

Abstract
This heuristic study is an attempt to introduce a nonparametric model to computerized adaptive testing in six steps. Elegance of the nonparametric model is that it put low restriction on data set and computations of its algorithms are so straightforward. In a two-dimensional data matrix a set of 30 multiple choices items were arranged by their difficulty order in total population. For each item there will be a column vector of row scores of length j. Then there will be i×j matrix formed by these items and scores. At every given row score, two ratios of right and wrong responses for each item were calculated singly, and then these two ratios were transferred to natural numbers. Using natural logarithm for these two ratios with exponents of total correct ratio U in numerator and incorrect total ratio W in denominator yielded a posterior estimation, OPT (Z) ≅0, for test items at every given row score.
For single items a Nonparametric Item Response Function (NIRF) were estimated over discrete row scores. At last, conditional error, lower limit, and upper limit of the test scores were estimated by a heuristic sub Bayesian theorem. Empirical data was used for checking method implementation. Results showed that the method will enable users to calibrate test items. Multidimensionality in data sets and Items with high guessing parameters, however, may cause some inconsistency in stochastic ordering of the latent trait by row score.

Evaluating model fit and skill interdependency of granular growth model in a formative assessment system

Author
Jinah Choi (Edmentum, Inc.)
Windy Torgerud (Edmentum, Inc.)

Abstract
Detailing students' academic growth has been one of the primary concerns in the field of education. As a part of procedures for developing granular growth model in a formative education system, this study performs model parameter estimation and model comparison via cross validation in RStan (an R package) in order to select models which best explain data. Namely, the model details probability of student mastery of skills encountered in a standard educational classroom environment through the use of a Cognitive Diagnostic Model framework, specifically deterministic inputs, noisy "and" gate (DINA) models.
Students encounter skills in an order determined via a "Learning Progression", a human expert generated sequence of skill development. Each item is written to assess a skill, and each skill belongs within a domain. Students navigate the learning environment first by completing a computer adaptive test (CAT), this test places students individually within the Learning Progression. Students encounter mastery tests and progress checks as they learn material. After some time of learning students complete another CAT and are placed into an updated sequence of skills based on their ability levels. Our models presented here take student performance data into account from their CAT, mastery test, and progress check item responses to inform an estimate of student probability of skill level mastery. We run the model on a subset of student data from early Fall semester to a subset of student data in late Spring to capture how student's probability of skill/domain level mastery changes over time, providing us a granular view of student growth.

Although the skills are learned in a linear sequence defined by the Learning Progression, the degree for which each skill depends on mastery of a previous is unknown. We run a model assuming skill independence and compare the model's performance (ability to predict left out data) via leave-one-out (loo) cross validation, to a model which is initialized with a Bayesian prior of some level of dependence of skills earlier in the learning progression to skills later. The model will learn via Bayesian update this skill dependency matrix.

Realigning the Scale of One Item Pool to Another: IRT Linking Using Growth Data

Author:
Yeow M Thum

Abstract

When well-designed, there are many advantages to computerized adaptive testing (CAT) over the conventional fixed-form test (Reckase, 1974; Green, Bock, Linn, & Reckase, 1984; Weiss, 2011; Weiss & Kingsbury, 1984). By constructing a unique test for each examinee, CAT scores are unbiased, more precise, and they are both more information efficient and effective for measuring examinees at all levels of a trait than fixed form tests. From the practical standpoint, CATs quickly return an examinee's score upon completion of the test, making test results readily available and therefore far more actionable for students, teachers, and schools.

But as Wise and Kingsbury (2000) have pointed out, even when well-designed, a CAT may be far more complex to operate than fixed-form tests. For example, many of the decisions involved in constructing a unique test for each potential examinee need a reliable software engineering support infrastructure. Maintaining the core functioning features of any assessment program may also be challenging. This study is set against the backdrop of a scale maintenance effort for a widely-used set of CAT assessments.

Routine maintenance of a computerized adaptive testing (CAT) program found that the scales of two CATs that measured an underlying developmental continuum were mis-aligned. In this presentation, we describe an approach to realigning the scales of the two CATs and discuss the results of this realignment. Realigning two CAT scales, like the linking or equating of fixed-form tests, concerns the determination of unknown form effects, a task that can be challenging when the banks contain no items in common or when the tests are not suitable for examinees at the same point in their development. Despite significant advances in test form linking or equating in many challenging situations, aligning or realigning the scale of one CAT item pool to another presents a unique problem (von Davier, Holland, & Thayer, 2004), particularly when items from one pool are not suitable for use in the other. Mis-alignment of the scales could be detrimental when using scores from both assessment as if they are to share the same equal-interval vertical scale that measures continuous development spanning kindergarten through middle school.

Drawing from recent proposals that employed ancillary information from examinees to construct pseudo-equivalent groups or to match examinees on their performance on the tests to be linked (Haberman, 2015; Longford, 2015), we employed growth model predictions using examinee longitudinal test results to provide common-person linking information that were otherwise missing by design. With the synthetic linking data, standard IRT linking approaches were then applied to rescale the item pools of one of the CAT scales. The approach proved effective

in ameliorating offending test-specific patterns of gains in pre-alignment scores. Content experts also found the re-calibrations of affected items reasonable. As a result of this adjustment, we believe that the representation of growth over the grades spanned by the tests is better justified.

Symposium: Converting from Paper-and-Pencil to CAT: Developing CAT Designs to Meet Required Constraints

Authors
Mark D. Reckase (Michigan State University)
Andrés Páez (Icfes)
Sewon Kim (Michigan State University)
Unhee Ju (Michigan State University)

Abstract
This session includes a series of presentations that show the process and results of a conversion from a high-stakes paper-and-pencil test to a CAT that meets all necessary constraints and improves the overall quality of the measurement. The example uses PreSaber 11 exam, that is an examination used by the Colombian Institute for the Educational Assessment and Evaluation (Icfes) as a mock exam that allows examinees to familiarize with the real high-stakes test. The latter is used for admission to undergraduate programs in Colombia. Because the CAT version of the examination were planned for using as practice tests, we did not have to take into account some of the components of a high-stakes examination such as the exposure control. However, the goal was to produce a CAT that could be modified for operational use in the future.
Session 1*: PreSaber 11 Design and Uses*
Andrés Páez, Icfes
The first presentation in this symposium describes the structure of the current PreSaber 11 testing program, how the program is used, and the design of the component parts. The PreSaber 11 program contains five separate tests, each of which has different specifications and constraints; that were developed using Evidence-centered Design, i.e. models based on evidence, competence, or task applied to the different subjects. For each of these components, important constraints were identified. The constraints formed basic design features for the corresponding CAT versions.
Session 2: *Designs of CAT Versions of the PreSaber 11 Components*
Sewon Kim, Michigan State University
For each of the five component parts of PreSaber 11, a unique design was developed for the corresponding CAT version. These different designs were needed because of the different constraints for each component. For example, some of the components had

stimulus sets with multiple test items that needed to be administered together. To meet these design requirements, hybrid CAT designs that combined features of item level adaptive tests and multi-stage tests were implemented.

Session 3: *Item Pool Designs and CAT Design Evaluations*

Unhee Ju, Michigan State University

For each PreSaber 11 component, initial design work was done to determine the characteristics of item pools needed to support the adaptive algorithm. The CAT designs were evaluated using simulated item pools to determine if they worked as expected. The CAT designs were also evaluated, using available item pools from Icfes, to determine if more item development was needed before the expected properties of the CAT designs could be achieved.

Session 4: *Final CAT Designs and their Properties*

Mark D. Reckase, Michigan State University

At the end of the process, each of the components of PreSaber 11 had a unique design and each design was evaluated to determine if it met the design goals. Each of the designs and the evaluation results are presented in this session. Lastly, the properties of each component in the system is described to show how well they perform.

Symposium: Learning meets Assessment: Adaptation and Personalization at Scale for Practitioners

1. Integrating Learning, Measurement, and Navigation: Introduction of a Recommendation and Diagnostic (RAD) API

Authors

Ada Woo ACTNext by ACT, Inc.)

Stephen T. Polyak ACTNext by ACT, Inc.)

Lu Ou (ACTNext by ACT, Inc.)

Abstract

The past few years have seen rapid advances in educational technology (EdTech). EdTech drives the development of learning and assessment systems (LAS) that have blurred the line between traditional learning and assessment. Students can now consume education materials, review lectures, and take tests on the same platform. The wide use of LAS at universities and K-12 institutions spurs efforts to extend the capabilities of LAS for a better learning experience.

As a universal scalable extension to existing LAS, ACTNext, the innovation and research unit of ACT, Inc. has developed a Recommendation and Diagnostic (RAD) API that personalizes learning based on diagnostics and augments existing LAS in a seamless way. The innovative service continuously tracks student's learning, diagnose mastery, and deliver personalized recommendations for learning. It monitors measurement

events in real time as students practice their skills. Cognitive Diagnostic Models (CDM) and the Elo rating algorithm are used to assess learner's abilities over time. These real-time analyses continuously update predictions of whether a student has mastered certain skills and diagnose the areas that need further review. Based on a student's skills and deficits, the API is then able to provide personalized open education resource recommendations through OpenEd. Learners may use the recommended online learning resources to practice the skills they have yet to master.
The API is robust, efficient, scalable, and extensible. RAD is built on a scalable, industry-accepted Amazon AWS serverless lambda architecture backed by high-performance DynamoDB data access. It allows implementation of various algorithms and models through a modular set of features. A solution is an ACT Software as a Service (SaaS) capability that can be applied to a range of subjects and systems.
The RAD API has already been integrated into ACT Academy, the freely available learning and test preparation platform for ACT. It leverages the ACT's Holistic Framework of Education and Work Readiness, a publicly available framework based on four broad domains: core academic skills, cross-cutting capabilities, behavioral skills, and education and career navigation skills. It aims to accompany a learner's practice and facilitates learning as a personalized instructor.
In this session, the authors will discuss the adaptive LAS framework on which the RAD API is built. They will explain the underlying models and algorithms used in the RAD API and share results from simulation studies that evaluate their strengths and limitations. The discussion will focus on the implementation of the API as an extension of LAS and the evaluation of its scalability.

2. The Effects of Adaptive Learning in a Massive Open Online Course on Learners' Skill Development

Authors
Yigal Rosen (ACTNext by ACT, Inc.)
Ilia Rushkin (Harvard University)

Abstract. Digital learning systems are considered adaptive when they can dynamically change the presentation of content to any user based on the user's individual record of interactions, as opposed to simply sending users into different versions of the course based on preexisting information such as user's demographic information, education level, or a test score. Conceptually, an adaptive learning system is a combination of two parts: an algorithm to dynamically assess each user's current profile (the current state of knowledge, but potentially also affective factors, such as frustration level), and, based on this, a recommendation engine to decide what the user should see next. In this way, the system seeks to optimize individual user experience, based on each user's prior actions, but also based on the actions of other users (e.g. to identify the course items that many others have found most useful in similar circumstances). Adaptive technologies build on decades of research in intelligent tutoring systems, psychometrics, cognitive learning theory, and data science. More specifically, Cognitive Tutors utilize knowledge tracing to track knowledge acquisition and provide tailored instruction, by tracking performance on individual production rules in a cognitive model. Extensions to this model have included estimating of the initial probability that the student knows a skill, estimating of the impact of help features on the probability of acquisition, and integrating with models of item difficulty. However, these approaches typically do not consider pacing and require significant content design workload in order to create learning and assessment content. These limitations are critical in large-scale MOOC context. Pioneer studies on adaptive technologies in MOOCs indicated both technical feasibility and the educational promise. Despite the promise of adaptive learning, there is a lack of evidence-based instructional design, transparency in many of the models and algorithms used to provide adaptive technology or a framework for rapid experimentation with different models. Harvard University partnered with Microsoft Learning to develop ALOSI (Adaptive Learning Open Source Initiative) provides open source adaptive learning technology and a common framework to

measure learning gains and learner behavior. The key insights gained from the modeling and analysis work enable us to address the development of evidence-based guidelines for the instructional design of future courses and provides insights into our understanding of how people learn effectively. ALOSI uses Bayesian Knowledge tracing to both develop a predictive model of skills mastery for the learner, and improve the predictive attributes associated with the content. The key features in ALOSI's current adaptive framework include knowledge tracing and recommendation engine, while user modeling, feedback, and recommendation of targeted learning materials are in development. The engine improves over time from the use of additional learner data and provides direct insights into the optimization processes (by contrast with commonly used commercial "black box" adaptive engines). Additionally, the architecture of the adaptive engine enables rapid experimentation with different recommendation strategies. This pilot study measured the effects of adaptive pathways on learning gains and dropout rates using different tuning parameters in the adaptive engine against the instructional design learning experience.

3.Building and Picking the Right Model for Learning and Assessment. Notes for Model Developers

Authors

Michael V. Yudelson (ACTNext by ACT, Inc.)

Abstract. Computer-based and computer-assisted educational systems have penetrated our daily lives. Delivery of one-size fits all content, assessment, or learning is no longer an option as it is an outdated approach. Multiple fields of study are focusing on approaches to efficiently represent student knowledge. The central issues for these approaches are the accuracy and validity of the estimates made.

Psychometricians and cognitive scientists over the years have come up with many approaches to represent student knowledge and the process if its acquisition. Many papers were published on the goodness of fit of those approaches as well as comparisons between the approaches. Model choice and model comparison, however, are not a matter of statistical approach alone. In psychometrics and adaptive learning, a lot of attention is given to loadings of model parameters and their interpretation and vetting.

We are considering a case for model design and model choice that cannot be classified as strictly assessment or strictly adaptive learning. It is not typical for assessment since the data is being collected for students who consume practice question items over an extended period of time. It is not typical for adaptive learning either since the student-skill density is rather sparse and lacks enough skill attribute repetition. We demonstrate how the consideration of traditional and not-so-traditional assessment and learning models could become challenging.

To provide a reference point for our method, we are applying the same model selection procedure to one of the largest openly available datasets from the area of intelligent tutoring systems – 2010 KDD Cup challenge data from Carnegie Learning, Inc.

Symposium: Adaptive Measurement of Change (AMC): Identifying Psychometrically Significant Change One Examinee at a Time

Authors:

David J. Weiss (University of Minnesota)

Jieun Lee (Pearson)

Chaitali Phadke (Scantron)

King Yiu Suen (University of Minnesota)
Chun Wang (University of Washington)
Matthew Finkelman (Tufts University)

Change in K-12 reading achievement on two occasions

Jieun Lee (Pearson)
David J. Weiss (University of Minnesota)

Abstract

Significance of individual change can be determined by utilizing hypothesis testing in AMC measurements taken between two occasions. Previous studies have shown that hypothesis testing methods with AMC resulted in desirable false positive and true positive rates compared to the confidence interval approach under various simulation designs (Finkelman et al., 2010; Lee and Weiss, 2014; Phadke, 2017). The current study applied four hypothesis testing methods in K-12 reading data measured by AMC. The percentage of significant change for each hypothesis testing method and agreement between the hypothesis testing methods were evaluated. A simulation study based on the same item bank and estimated θs from the reading data also evaluated true positive and false positive rates of the hypothesis testing methods and comparisons were drawn between the results of the simulations and those from the real reading data.

Application of Omnibus Hypothesis Tests to K-12 Math Data in Measuring Growth

Chaitali Phadke (Scantron)
David J. Weiss (University of Minnesota)

Abstract

The present study applied six new omnibus hypothesis tests to measure psychometric significance of individual change. The omnibus hypothesis tests have shown a promising potential for identifying significant individual change when examinees are measured over multiple occasions by displaying an optimum balance of false positive rates and true positive rates (Phadke, 2017). Contrasts allowed determination of the occasions accounting for significant change when the omnibus changes were significant. The tests were applied to K-12 math data, obtained by taking measurements in the beginning, middle and at the end of the school year using AMC. Omnibus tests detected linear and non-linear change patterns with proportions of agreement between hypothesis tests ranging from 0.81 to 0.98. The hypothesis tests displayed similar detection rates for live AMC data as for simulated data (Phadke, 2017).

Change in multiple patient-reported outcomes across multiple occasions

King Yiu Suen (University of Minnesota)
Chun Wang (University of Washington)
David J. Weiss (University of Minnesota)

Abstract

In psychology and educational measurement, it is often of interest to assess change in an individual. The current study expanded on previous research by introducing methods that can evaluate individual change on multiple latent traits measured on multiple occasions. Simulation

studies were conducted to examine the true positive rate and the false positive rate of the new methods under a conventional test and a computerized adaptive test. Manipulated variables included the number of occasions, change magnitudes, patterns of change, and correlations between latent traits. Real-data examples are provided of measuring and identifying psychometrically significant change across two to four occasions using a multivariate self-report medical outcome measure from hospitalized patients.

Time efficient adaptive measurement of change

Matthew Finkelman (Tufts University)
Chun Wang (University of Washington)

Abstract

AMC uses CAT at multiple occasions to efficiently assess a respondent's improvement, decline, or stability from occasion to occasion. Whereas previous AMC research focused on administering the most informative item to a respondent at each stage of testing, the current research proposed the use of Fisher information per time unit as an item selection procedure for AMC. The latter procedure incorporates not only the amount of information provided by a given item, but also the expected amount of time required to complete it. In a simulation study, the use of Fisher information per time unit item selection resulted in a lower false positive rate in the majority of conditions studied, and higher true positive rate in all conditions studied, compared to item selection via Fisher information without accounting for the expected time taken. Future directions of research are suggested.

An Evaluation of Collapsing Response Categories for Innovative Question Types Using the GPCM

Author

Darrin M. Grelle, SHL

Abstract

For several decades, cognitive ability tests used for pre-employment selection have been primarily dichotomously scored multiple choice tests. In the current economic climate, the hiring process has become candidate-centric and organizations are looking for mobile compatible tests that candidates find more engaging and job relevant than traditional cognitive. This study centers on a numerical reasoning test created specifically to meet these market demands and constructed utilizing mobile first responsive web design. Test questions offer candidates the ability to solve engaging, job relevant scenarios through various interactions on screen (e.g. adjusting wedges of a pie chart or ranking avatars based on a set of rules). This test is adaptive, with the goal of creating the most concise and accurate test possible. Each question has multiple response points that lend themselves to the generalized partial credit scoring model (GPCM; Muraki, 1992).

The challenge in implementing these innovative and interactive item formats within an adaptive test, powered by the GPCM, was determining how to score the content before item parameters could be estimated. The variety of interactions designed into the questions led to a variety of uncertainties related to scoring such as what to do with:

- missing response categories due to lack of endorsement, which has been shown to be an issue (Wilson & Masters, 1993),

- items with response categories that are missing by design (e.g. in a pie chart question with four wedges, it is impossible to get only three correct) or due to lack of data,.
- items that require candidates to place objects in a sequence where they could be scored by whether each object was placed in the correct slot, or if each object immediately followed the correct object.

In order to address the scoring challenges, test questions were trialed on a sample of 14,729 examinees drawn from an online test prep site and a crowdsourcing site. Each examinee responded to 12 questions with an 18 minute timer. The 12 questions were randomly drawn from a bank of 279 questions. The number of "points" possible for each question varied from one to six (Mean: 3.46). Parameters were estimated for several versions of the data using the MIRT package in R. A 2x4 design was used to generate eight datasets for parameter estimation. The first condition was whether sequence questions were scored by correct slot or correct preceding object. The second condition was the degree to which categories were collapsed. Categories were either not collapsed at all, collapsed only for categories with n=0, collapsed for categories with fewer than 5% of the total responses for that question, or collapsed to dichotomously scored. The theta estimates and standard errors will be shared for each condition. Our results showed lower standard errors and better construct validity when response categories were collapsed for zero and low volume categories when sequence questions were scored by correct slot. We believe this research will inform those working towards the implementation of CAT using innovative item types with nontraditional response formats.

Impact of Innovative Item Scoring on Constrained Computerized Adaptive Testing

Authors
Xin Lucy Liu (Senior Psychometrician, Ascend Learning)
Xuechun Zhou (Senior Psychometrician, NCS Pearson)
Haiqin Chen (Psychometrician, American Dental Association)

Abstract
Innovative item types are increasingly being used in assessments due to their capability of effectively assessing higher-order skills that cannot be measured well using traditional multiple-choice items (e.g., Huff & Sireci, 2001; Sireci & Zenisky, 2006). Despite the great promise of the measurement quality, innovative items are still scored using dichotomous scoring algorithm in many operational computerized adaptive testing (CAT). The studies on the CAT implementations of such items are limited in an operational setting. Some of the few related studies cannot fully examine the impact of polytomous scoring of these items on CAT mechanism due to practical limitations (e.g., Jiao, Liu, Haynie, Woo, & Gorham, 2012).
The purpose of this study is to investigate the impact of different scoring methods of innovative items on the measurement accuracy and item pool utilization for a constrained CAT program. Two types of innovative items types, multiple and ordered response, are considered for this study. Two scoring algorithms of these items, dichotomous and partial credit scoring, are examined. The CAT program is modeled after a high-stake mixed-format operational CBT program that includes both traditional multiple-choice items and innovative items.

The Influence of Dimensionality on Item Exposure Rate under CAT Administration

Author
Mingjia Ma
Terry Ackerman

Abstract
Test security is an important issue in the administration of a computerized adaptive test (CAT). An over-exposed item is likely seen by an examinee before the test and possibly brings an unfair advantage to the examinee. Thus, test developers try to control item exposure rate under different conditions in order to protect the integrity of the test. Such conditions include a pool being multidimensional.
Under the CAT environment, examinees receive different items from the pool based on their ability estimates, so examinees will receive different test forms. Because of this, dimensionality of computerized adaptive testing could be a concern of the CAT item pool development. If the item pool is strictly unidimensional, there will be no concern of dimensionality (Wang & Kolen, 2001) and the same measured latent trait will represent the same ability for every examinee even though they received different test forms. However, due to different item pool building process, the item pool may represent a combination of several sub-domains of a subject area and it's unlikely that the item pool is unidimensional. Under such consideration, researchers have proposed different methods to combine the correlated dimensions. Ip (2010) proposed the Projective IRT approach to get rid of the invalid dimension of test design; Camilli (1992) proposed the reference composite approach to mathematically derive the unidimensional item parameters from multidimensional latent space. Both approaches could transform the multidimensional property of an item pool to unidimensional, but we do not know which one could help reduce the item pool exposure rate when the pool is applied to CAT administration. Also, with different level of latent trait correlation, the optimal way of getting the unidimensional latent trait varies.
This paper tries to compare these two approaches and to investigate which one could result in a better exposure rate control at different levels of latent trait correlation.

Conditional Exposure Control in Adaptive Testing

Authors
Hao Ren (Pearson)
Qi Diao (ETS)

Abstract
One of the important components of an operational computerized adaptive testing (CAT) program is the exposure control method. Conditional exposure control methods (Stocking & Lewis, 1998; van der Linden & Veldkamp, 2007) are designed to be more effective in reducing the

likelihood of test takers of similar ability level being exposed to the subset of items in the pool specific to them. The dilemma of conditional exposure control methods is that in real world, the true ability of the test taker is unknown while the methods are trying to control the exposure of test takers of similar ability levels. A consequence of such dilemma is that the goal exposure rates may not be achieved based on test takers' true abilities by using traditional conditional exposure control methods. Even though we cannot fully control based on true abilities, this study compared three methods that were designed to improve the performance of conditional exposure control methods in adaptive testing. The first method was based on controlling maximum exposure rates. The second method was based on combining marginal and conditional exposure control methods. And the third one was based on adding random draws on the ability bands. The three methods will be discussed and the results of the simulations will be presented to show that even though we cannot fully control the exposure based on students' true abilities, the methods described in the study can improve the performance of conditional exposure in adaptive testing.

Detecting Aberrant Behavior in Computerized Adaptive Testing: The Lognormal Response Time Model

Author
Xiaowen Liu
H. Jane Rogers

Abstract
Most states use a statewide assessment system to describe student achievement and growth of student learning as part of program evaluation and school, district, and accountability systems. Some states, such as Connecticut, use Smarter Balanced Assessments to measure student progress in English language arts and mathematics. However, these assessments often do not have significant consequence for students. Examinees with low motivation may not give their best effort and thus show aberrant behavior in their responses. Therefore, students' aberrant response patterns need to be monitored, flagged, and explored to ensure the validity of results obtained from these testing programs. Procedures for the detection of aberrant responses are well developed. With the advent of computerized test administration, information about response time is also available and may be useful in detecting aberrant response behavior. Response time (RT) is an indicator that reflects information about respondents' speed and mental activities, as well as item and test characteristics. Recently, Marianti et al. (2014) applied the lognormal response time model and proposed a likelihood-based person-fit statistic to detect different types of aberrant test-taker response times in fixed form tests. This procedure could be used to flag respondents and items for further consideration. The current study aims to adjust the person-fit statistic proposed by Marianti et al. (2014) to investigate its performance in detecting different types of aberrant responding in the context of computer adaptive testing.
A simulation study is designed to test the performance of the person-fit statistic in the context of computerized adaptive testing. Three types of aberrant response times were simulated in the dataset: random, fast, and extreme response times. For the three types of aberrant responses, three different proportions of test-takers with aberrant response times were generated: 5%, 10%, and 20%. The sample size was 1000

examinees and the test length is 40 items. For each student, 40 items were randomly drawn from an 80-item bank. R package LNIRT would be used in fitting the lognormal response time model and computing the person-fit statistics. Fifty replications were performed for each condition. The percent of aberrant responders correctly identified (hit rate) and the percent of non-aberrant responders incorrectly identified (false alarm rate) would be computed in each condition. Additionally, RMSE and Bias indices were computed for person and item response time parameters. Results showed the condition with random aberrant behaviors had high detection rates.

Effects of Compromised Items in Computer-Adaptive Tests: A Retrospective Study

Authors
Luz Bay

Abstract
Test security is one of the most touted advantages of a computer-adaptive test (CAT) when compared to a paper-and-pencil test. The reason for which is the small likelihood that when a student somehow divulges the test items they took, another student taking the test will be presented with the same set of items, thus having an advantage and receiving a higher score relative to their actual ability. And yet for an assessment program that is computer-adaptive, a very conservative practice of deactivating items when there is reason to believe that they have been shared has always been observed. Such practice is observed to assure constituencies of the very secure nature of the assessment, assuring the public that the validity of the test is not threatened. With the assessment program scheduled for a sunset date, we are studying in retrospect whether the rigmarole that goes with removing compromised items from the item pool, with all the resources it requires, was warranted. Using data from the last full year of two assessments, percentages of item overlaps with set of items for randomly selected students were computed. We also studied how much exposure an item string needs to have before it has material effects on the average score of the student population, with variation on the ability level associated with the item string.

Using Response Time to Improve Precision and Efficiency of Computerized Adaptive Testing

Author
Jing Lu
Chun Wang (University of Washington)
David Weiss (University of Minnesota)

Abstract

Fan, Wang, Chang, and Douglas (2012) proposed a time efficient item selection method in CAT, namely, the maximum Fisher information per time unit (MFT). The primary idea is to maximize the amount of information accrued per time unit, so that the testing time could be reduced without sacrificing measurement precision. In this study, we extend their method in two aspects. First, in addition to use response time (RT) during item selection, the RT is also used as collateral information during interim latent trait update. Hence, it is expected that not only the testing time will be reduced, but also the measurement precision will be further enhanced. Second, we design the study using polytomous items from the graded response model, which is suitable for general Likert scale items. 324 items from the Activity Measure for Post-Acute Care (AM-PAC) calibrated using the bivariate joint model (Wang, Weiss, & Su, 2019) are used in the study. The proposed item selection and interim update method will be compared to two reference methods, the original MFT method and the simple maximum Fisher information method, in a variable length CAT. It is anticipated that the new method will greatly reduce the average test length.

Chance-constrained test assembly

Authors
Giada Spaccapanico Proietti (Department of Statistical Sciences, University of Bologna, Italy)
Mariagiulia Matteucci (Department of Statistical Sciences, University of Bologna, Italy)
Stefania Mignani (Department of Statistical Sciences, University of Bologna, Italy)
Bernard Veldkamp (PhD: Faculty of Behavioral, Management and Social Sciences, University of Twente, The Netherlands)
Angela Verschoor (Cito, Arnhem, Netherlands)

Abstract
The recent developments in computer technology enabled test institutes to improve the test assembly process by automated test assembly (ATA). A general framework for ATA consists in adopting mixed-integer programming models. These models are intended to be solved by common commercial solvers which, notwithstanding their success in handling most of the basic ATA problems, are not always able to find solutions for highly constrained and large-sized instances such as the assembly of several parallel test forms. Moreover, all parameters are assumed to be fixed and known, an hypothesis that is not true for estimates of item response theory (IRT) parameters.
These restrictions motivated us to find an alternative way to specify and solve ATA models. First, we suggest a *chance-constrained* approach (CC), see Charner (1959) for a detailed introduction, which allows to maximize the *(1-alpha)*-quantile (usually greater than 0.90) of the empirical distribution function of the test information function (TIF) obtained by bootstrapping the calibration process. Secondly, for solving the ATA models, CC or not, we apply a stochastic heuristic called *simulated annealing* (SA) proposed by Goffe (1996). This technique can handle large-scale models and non-linear functions of which the chance constraints are an example. A Lagrangian relaxation approach helps to find the most feasible/optimal solution and, thanks to a random variable, more than one neighborhood of the space is explored avoiding to being trapped in a local optimum. Several simulations are performed and the solutions are compared to the results of CPLEX 12.8.0 Optimizer. All the described algorithms are coded in the open-source framework Julia.

A tree-based algorithm for test blueprint creation

Author
Tyler Matta

Abstract
Prior to the advent of computer adaptive testing, building personalized assessments for individual students was a laborious task rarely taken on by educators. Today, adaptive test algorithms have enabled educators to administer personalize assessments based on student's unique location on a given construct. Because the underlying blueprint used to drive that adaptive algorithm is designed by assessment experts, most adaptive assessment products employ a common blueprint such that each student, regardless of time and place, will be administered an assessment that corresponds to the same design. So, while adaptive assessments are highly efficient at personalized item delivery, they remain crude with regards to the unique design requirements of educators.
This paper presents a tree-based algorithm for test blueprint creation. The algorithm design draws on the premise that many constructs targeted by educational assessments are composites of a (potential large) set of knowledge, skills and abilities, referred to generally as states. The mapping of certain composite constructs to the set of states embodies a tree-based structure.
The first stage of the test blueprint algorithm is to create regularized weights for each node at each level of the hierarchical construct map. The weights are based on leaf size, which is a relative property of each state. The second stage uses these weights with additional global test constraints (e.g., total test length) to determine which level of the hierarchy to place constraints, and their appropriate size. There are multiple scenarios when the algorithm may not find a blueprint that satisfies the global constrains. In such cases, either deterministic or probabilistic constraints may be used to satisfy aspects of the design, enabling the algorithm to work through the remainder of the test blueprint.
In addition to presenting a technical description of the algorithm, this paper demonstrates how the algorithm can be initialized by a simple user-interface. Next, the paper goes on to show how the test blueprint algorithm generalizes to multidimensional contexts. Finally, the paper demonstrates applicability with both standards-based assessments and assessment targeting constructs with an underlying learning progression.

A Search for a Practical Definition of an Optimum Item Pool

Author
Emre Gonulates

Abstract
Item pools are a crucial component of adaptive tests. Ideally, adaptive tests use optimum item pools, which is defined in this study as an item pool presenting an examinee an optimum item regardless of the stage of the test and the ability level of the examinee. An optimum item, for an ability level, is defined as an item that provides the maximum possible information at that ability level.
However, in practice, it is not possible to create such an optimum item pool because an optimum item cannot be created (for certain measurement models).
The problem with the optimum item is that, theoretically, for measurement models with item discrimination parameter, it can provide infinite amount of information at a given ability level. Such a definition for an optimum item is not useful in practice due to its elusiveness.
Even if an optimum item is created, to create an optimum item pool, not just many but infinitely many of optimum items are necessary.
In this study, the theoretical definitions of an optimum item and item pool will be presented. Additionally, the difficulties of using these definitions in practice will be discussed. Then, practical limits of an optimum item and optimum item pool will be discussed. The first argument will be about the meaning of very high item discrimination from a substantive point of view, specifically for achievement tests. Second argument will be from a statistical point of view: (1) under which conditions high item discrimination parameters are observed, and, (2) the limits of item discrimination and item information in practical settings. Finally, an optimum item pool definition for practical settings will be proposed.
Since the adaptiveness of an adaptive test is directly tied to the quality of an item pool, comparing an item pool with an optimum item pool will allow researchers to evaluate the quality of their item pools.
The knowledge of the limits of the optimum item pool will enable test developers to quantify the degree to which their item pools deviate from the optimum item pool. This will lead better evaluation of the quality of the adaptive tests.

On Measuring Adaptivity of an Adaptive Test

Author
Zhongmin Cui (ACT, Inc.)

Abstract
Reckase, Ju, and Kim (2019) raised an important question: “How adaptive is an adaptive test?” Although many tests are labeled as computerized adaptive test (CAT), not all tests show the same degree of adaptivity – some tests might not have much of adaptation because of various constraints imposed by test developers.
Reckase et al. proposed three indices to measure the amount of adaption for an adaptive test: (1) the correlation between the mean *b* parameter values and the final theta estimations for examinees; (2) the ratio of standard deviations between the mean *b* parameter values and

the final theta estimations for examinees; and (3) the proportion of reduction of variance of *b* parameters for the items selected for examinees compared to the variance of *b* parameters for all items in the bank. “These statistical descriptors provide slightly different information about the amount of adaption”, they argued, “[t]ogether, they provide objective information about the matching of the item selection to the level of performance for everyone taking the test.”
These statistical descriptors, however, may provide misleading information on the amount of adaption – the author will present a simulation study to show the problem. Also in the presentation, a new index of adaptivity will be proposed. The new index will be compared with the aforementioned three indices in a simulation study under various conditions (e.g., different bank sizes, different variances of b parameters, and different CAT administrations). The full details of the simulation procedure and results of the simulation study will be included in the final paper. The author expects this study will provide guidelines for practitioners in measuring the adaption of CAT and inspire more research in this area.

A Framework for Measuring the Amount of Adaptation of Rasch-based CATs

Authors
Adam E. Wyse
James R. McBride

Abstract
A key consideration when giving any computerized adaptive test (CAT) is how much adaptation is present when the test is used in practice. A test with a high level of adaptation is one where the items selected for an examinee are well watched to their ability level. There are several factors that can impact how adaptive a CAT is when it is given to an examinee, including the size and quality of the item pool, the constraints used in the CAT, and the ability level of the examinee. Previous work by Reckase, Ju, and Kim (2018) suggested that one can measure the amount of adaptation of a CAT by looking at how well the items administered to an examinee match the examinee’s overall ability level at the end of the test. Under this framework, a test is considered to be highly adaptive when the items given in the test well match the examinee’s final ability estimate. However, during an operational CAT the examinee’s final ability level is not known and items are selected to match provisional estimates of ability at the start of each question. The basic principle of the new framework is to measure the amount of adaptation of a Rasch-based CAT by looking at the differences between the target item locations for an examinee based on their estimated ability level at the start of each question and the selected item locations of the items they end up getting administered. In this framework, a highly adaptive test is one in which these differences are small for examinees. Several numerical and graphical methods based on the new framework are illustrated and discussed. These numerical and graphical methods are compared to the measures of adaptation suggested by Reckase et al. (2018) using simulated data and data from an operational K-12 Rasch-based CAT. Results suggest that the new methods can effectively measure the amount of adaptation in Rasch-based CATs and can provide important information regarding situations in which the CAT may not be highly adaptive for individual examinees or groups of examinees. It is suggested that some of the numerical and graphical methods introduced may provide better measures of the amount of adaptation of a Rasch-

based CAT than the indices suggested by Reckase et al. (2018). Discussion is given to how the numerical and graphical methods may be applied to evaluate the level of adaptation of other Rasch-based CATs.

Three Measures of Test Adaptation Based on Optimal Test Information

Authors
G. Gage Kingsbury
Steven L. Wise

Abstract
This study extends the work of Reckase, Zu, and Kim (2019) by introducing three new measures of test adaptation. Reckase et al. identified that many approaches to testing are called adaptive testing today, but the many constraints of content, test design, item pool usage, and item selection make them vary widely in the amount of adaptation that is actually accomplished for a group of test takers. The idea behind the work of these authors was to establish metrics that would allow differences among the adaptivity of different tests to be established and quantified.
The current study develops a new class of adaptivity measures that extend their utility. In addition to quantifying differences in adaptivity for groups, they are designed to examine theoretical, operational, and engineering questions concerning adaptive tests, so that statements about their quality can be made, and so that avenues can be identified to improve the adaptivity of tests for groups or subgroups of individuals in the testing population.
These measures are based on the amount of information that the test provides to individual test takers and comparing that value to an important criterion value (either the maximum information possible, the maximum information available in the current item pool, or the maximum information that could be obtained at the observed trait level estimate). Since they are information-based metrics, their utility in identifying the impact of the quality of adaptation on the observed trait level estimates should be fairly direct, and their utility for engineering better item pools and test designs should also be straightforward.
In this study, the three measures of adaptation are applied to simulations of adaptive, multi-stage and fixed-form tests to describe how the measures could be used in an operational setting to identify difficulties with adaptation, and to suggest processes for improving adaptivity. In this simulation, the same item pools and item characteristics were used to create three different tests, using traditional test designs. The simulated test takers then generated responses to all items in the item pool, and those responses were used to provide each sim with three testing outcomes for the three test designs. The outcomes were then used to illustrate the characteristics of the three new measures of adaptation, as well as those described by Reckase et al. (2019).
In addition to the overall analysis (which replicates the findings of the earlier authors) examples are given to show how the new metrics extend the utility of the technique into one that can be used to compare performance of test designs for subgroups of test takers, to the comparison of different constraints within a test design, and to the potential improvement of adaptivity for populations or for subgroups.

The need for adaptivity measures is obvious in a world in which many are trying to implement adaptive testing without advanced knowledge of the psychometric impact of their design decisions. Using the new metrics to provide practitioners with data that can help them improve their item pools and test designs may point them toward important considerations.

10/66 Dementia Research Group Cognitive Test Battery: results from IRT and CAT-simulation studies.

Authors
Fernando Austria (National Institute for Educational Assessment, Mexico)
Nate Thompson (Assessment System Corporation, Inc, United States)
Dafne Ortiz-Saavedra (National Autonomous University of Mexico, Faculty of Psychology)
Astudillo-García, C.I. (Psychiatric Attention Services, Health Services, Mexico)
Acosta-Castillo, I.G. (Dementia Laboratory, National Institute of Neurology and Neurosurgery, Mexico)
Sosa-Ortíz, A.L (Dementia Laboratory, National Institute of Neurology and Neurosurgery, Mexico)

Abstract
The purpose of this study is to develop a Computerised Adaptive version of the cognitive test battery used in 10/66 Dementia Research Group, a nationwide research project in Mexico. The battery currently consists of 30 dichotomous items and 5 polytomous items; the goal is to develop a CAT version using dichotomous items that accurately recaptures scores from the full assessment while using far fewer items, increasing its utility as a nationwide screening instrument for healthcare professionals.
The study utilized secondary analysis of the basal evaluation of the Dementia Research Study 10/66 in Mexico (n=2003). The 10/66 Dementia Research Group cognitive test battery included the administration of the Community Screening Instrument for Dementia (CSI 'D'), developed by the Consortium to Establish a Registry for Alzheimer's Disease (CERAD), with component tests for verbal fluency (VF), word list memory (WLM, immediate recall), and recall (WLR, delayed recall). For the purpose of this work, only 29 items from the CSI'D' were used; one was removed for poor fit. The assessment utilizes a cutscore of 0.65, which was set by fitting machine learning models to predict actual physician diagnosis using scores from the battery.
Item calibration was performed in the Xcalibre 4.2 (Guyer & Thompson, 2014) software using a two-parameter model. Computer Adaptive Testing was then simulated using the CATsim and SimulCAT applications to assess item bank performance through two different algorithms for item selection and termination criteria.

Two-parameter model analysis showed an excellent fit to the model in 29 items from 30 ($p>.05$), and the assessment had sufficient reliability given its short length (Cronbach's $\alpha > .85$). Analysis also provided evidence that the assessment was unidimensional and free from local dependency.

In the first simulation study we used SimulCAT (Han, 2012) with the item selection criterion of Efficiency Balance Information & no exposure control, and EAP score estimation with the initial score value chosen between -1 to +1, and the termination criteria when SEM became smaller than 0.600, minimum 5 items, maximum 15 items. The item bank was successfully administered three replications using an average of 14 items (12 for positive cases, and 15 for non-positive cases). EAP score estimations by trail did not differ significantly from the true theta ($F=.815$, $p=.565$), while there was a significant difference between by diagnoses as expected ($F=141.08$, $p=.001$). Post-hoc Scheffe test confirmed that there was not a significant difference between the simulations and the real application.

In the second simulation study we used CATsim (Weiss & Guyer, 2010), with the item selection criterion of Maximum Fisher Information, no exposure control, maximum likelihood score estimation, the initial score values at 0.0, and the termination criteria varied from 0.6 to 1.0, with an optional a maximum of 20 items. Results were again promising. The condition with SEM=0.8 and a 20 item maximum produced scores that agreed with diagnosis on the linear form 97.5% of the time, but with only 10.672 items on average; a drastic reduction from the original 30-item linear form.

The current research demonstrates that brief 10/66 CSI'D' item bank can support fundamental measurement and perform well in simulated CAT situations. The results suggest that the CAT version can produce accuracy classifications with far fewer items, providing a very effective yet short tool for healthcare professionals to use in screening patients. However, more research is required to optimize the functioning of the algorithm by age and educational level that are variables that potentially covariate with cognitive decline.

Application of CAT Stopping Rules to Increase Precision and Decrease Burden for Health Outcomes Research

Authors

Kathryn L Jackson, MS (Northwestern University, Department of Medical Social Sciences, Chicago, IL)
Ben Schalet, PhD (Northwestern University, Department of Medical Social Sciences, Chicago, IL)
Michael Kallen, PhD (Northwestern University, Department of Medical Social Sciences, Chicago, IL)
Aaron Kaat, PhD (Northwestern University, Department of Medical Social Sciences, Chicago, IL)
Michael Bass, MS (Northwestern University, Department of Medical Social Sciences, Chicago, IL)
Richard Gershon, PhD (Northwestern University, Department of Medical Social Sciences, Chicago, IL)
David Cella, PhD (Northwestern University, Department of Medical Social Sciences, Chicago, IL)

Abstract
Computer Adaptive Tests (CATs) are increasingly becoming a conventional way to measure health outcomes. The National Institute of Health Patient-Reported Outcome Measurement Information System (PROMIS) measures allow for measurement of multiple patient reported outcomes across chronic health conditions. PROMIS measures are based on banks of items calibrated using item response theory, and can be administered as CATs; final scores are presented as t-scores with a mean of 50 representing the general population, and standard deviation (SD) of 10.
Default PROMIS CAT administration stops after a minimum of four items and either a standard error (SE) of 0.3 has been reached or a maximum of 12 items has been administered. Two suggested improvements to administration are to reduce the number of items administered (reducing patient burden), and to increase score precision. Decreasing the acceptable SE for stopping administration will, in turn, increase reliability, correlating to an increase in precision (decrease in MSE) for the test. This, however, also causes an increase in number of items administered. To mitigate the increase in items, we suggest an addition to the CAT stopping rules which would stop administration if the reduction (or predicted reduction) in SE is negligible. Here, we focus on the application of these rules to increase precision (decrease MSE), while identifying the cost in patient burden (CAT length).

Examining Exposure Control on the NIH Toolbox Picture Vocabulary Test

Authors
Aaron J Kaat ((Northwestern University Department of Medical Social Sciences)
Richard C Gershon (Northwestern University Department of Medical Social Sciences)

Abstract
The NIH Toolbox Picture Vocabulary Test v2.0 (TPVT) is a currently-operational computer adaptive test (CAT). The TPVT includes 342 items, which were developed and calibrated using the Rasch/one-parameter logistic model. It has utilized two different randomization strategies for CAT-based exposure control and implemented them on different testing platforms, including selecting a random item within a predefined logit range around the current ability estimate, and selecting a random item from the top-N most informative items at the current ability estimate. However, as within-person TPVT assessments increased, concerns were raised regarding the potential failure of the exposure control mechanisms both within- and across-examinees.

Investigating Item Parameter Drift As a Problem of Ability Parameter Drift in CAT

Author
Beyza Aksu Dunya

Abstract
This simulation study was conducted to analyze impact of item parameter drift (IPD) on person ability estimation and classification accuracy when drift affects an examinee sub-group. The vast majority of IPD sources related to the factors that result changes in person ability. Those factors include differential opportunity to learn content of the test (Han & Guo, 2011). As such, we created CAT conditions where a sub-group of examinees' ability changed due to differential learning opportunities. Four factors were manipulated: (a) percentage of IPD items in the CAT item pool, (b) percentage of examinees affected by IPD, (c) ability parameter distribution of examinees and (d) direction of drift. The percentage of IPD items in the item pool varied by three: 20%, 40%, and 60%. The percentage of examinees who were affected by IPD varied by two: 30%, and 60% of the examinee sample. In some particular tests, such as certification and licensure exams, ability distributions are often skewed (Witt et. al, 2003). This can apply some of the educational testing situations too if majority of examinees in the sample perform higher or lower than the norm due to changes in learning opportunities. Thus, third independent variable was ability parameter distribution, varied by three levels: positively-skewed, negatively-skewed and normal distribution. Lastly, direction of drift changed by two levels: all IPD items drifted in negative direction, and all IPD items drifted in positive direction. Magnitude of drift ranged from 0.50 to 1.00 logits from their original values. Item parameters were drawn from a normal distribution. The test was a variable-length test and terminated when 95% confidence in the performance level could be made. 100 replications were generated for each condition. One-parameter item response theory model was used to calibrate items and maximum information criteria was employed for item selection. IPD impact on ability estimation was evaluated using multiple indicators; bias, root mean square error (RMSE), and mean absolute differences (MAD). A multivariate general linear model treating four factors as fixed effects and bias and RMSE as outcome variables are used to investigate both main and interaction effects. The impact of drift on classification accuracy was examined based on number and percentages of misclassifications, their significance, and rank order correlations. The findings revealed that IPD exposed to a sub-group of examinees can affect classification accuracy of those examinees substantially, but IPD impact on average ability estimation was small. The interaction between ability parameters distribution and drift direction was found significant. Most misclassifications occurred around the cut scores by the medium ability group of examinees for all distribution types. Thus, an important finding of this study was that testing companies should focus on item maintenance, particularly those with difficulties around the cut scores. This study provides useful information to schools, states, and countries planning to implement or are currently implementing CAT as part of their assessments by emphasizing IPD in sub-groups may violate fairness, invariance and consistency principles of CAT applications.

Exploring Differential Item Functioning in Cognitive Diagnostic Computer Adaptive Testing

Authors
Nixi Wang
Chun Wang

Abstract

Cognitive diagnostic computer adaptive testing (CD-CAT) has been an instrumental advancement in providing tailored diagnostic information and responsive estimation for examinee's latent traits. As part of the validity argument, the aspect and impact of differential item functioning (DIF) in CD-CAT needs to be investigated and addressed. More research needs to review the procedures of CD-CAT and methods of detecting uniform and nonuniform DIF in the CD-CAT framework. A design of a simulation study is proposed, requiring further steps for the performances and comparisons of different DIF procedures and various impacts in the context of CD-CAT.

There are many aspects of considerations that call for this research on DIF in CD-CAT to be investigated. First, though the development of CAT field and generating new methods is underway, one needs to pay careful attention about the validity of this test instrumentality. DIF could lead to invalid estimates and increased measurement error, while contributing to the test unfairness for examinees from different groups (Lord, 1980). Thus, potential threat engendered by DIF items could also complicate the CD-CAT operational situations. Finding a simulation study design mimicking the realistic calibration scenario is needed. It is yet unknown how and to what extent can CD-CAT self-adjust to DIF presence and types, and what DIF detection methods can be effective in what stages of CD-CAT. In the literature, sparse research is conducted for DIF situations in an administrative CAT scenario, such as Feng (2004) and Piromsombat (2014) looking at DIF impacts in operational CAT items. Other studies were mostly on precalibration process. However, those studies are based on IRT models, in the CD-CAT, the items exhibit DIF are conditioning on the attribute mastery profiles from different groups. No extant research has explored the DIF magnitude, location, and impact in the context of CD-CAT. In review of the finer grained meanings of DIF in person latent attributes that DINA model could illustrate, I propose a simulation design to investigate the effect of DIF in CD-CAT.

Application of Computerized Adaptive Testing using Medical Class Examination in South Korea.

Author
Dong Gi Seo
Sun Huh

Abstract
We have developed a platform to estimate a students' ability in order to provide freely available access to a web-based computerized adaptive testing (CAT) platform. We have used PHP and Java Script as the program languages, PostgresSQL as the database management system on an Apache web server and Linux as the operating system. A CAT platform allows for an administrator to input and search within the item bank and to construct tests. The CAT platform can estimate students' ability on each test based on a Rasch model and 2- or 3-parametric logistic models. Our CAT platform provides an algorithm for a web-based CAT, replacing previous personal computer-based ones, and makes it possible to estimate an examinee's ability immediately at the end of test. This study described the application of CAT platform and showed the result of CAT performance using medical class examination in South Korea.

The advantage of the CAT platform is web-based and allows for the immediate scoring after the termination of test. In addition, the selection of exposed items can be controlled because the exposure rates are recorded in the CAT platform. Practical issues in developing and maintaining a CAT program includes an item bank, test administration, test security issues. Out of these, item bank and test administration was addressed in our CAT platform. Test security and examinee issues should be dealt according to the policy by each institute during the implementation of the CAT platform. The CAT platform was not a freely available in South Korea certification/licensing examination, our CAT platform may open the horizon for an easily applicable web-based CAT in South Korea. It may be able to be used widely in the near future South Korea. This CAT platform may not be complete so that our CAT platform will be improved in the near future.

MYMAP: A Graph Theory Approach to Modelling Attribute Mastery

Author
Stephanie Varga

Abstract
MYMAP is a novel computerized diagnostic assessment for academics (CDA) that estimates mastery of an attribute hierarchy using a directed graph (digraph) and traditional computer adaptive testing (CAT). Two well-known theories that have proved useful for educational diagnostic testing are the Rule Space Model (RSM) and Partially Ordered Set Theory (POSET). Two computerized implementations of these theories are the RSM-CAT and the POSET-CAT. In this study, a Monte Carlo simulation of student responses to mathematics questions is performed, using item bank data taken from a comparison study of RSM-CAT and POSET-CAT. Neither RSM-CAT nor POSET-CAT uses a traditional CAT based on the item response theory (IRT) framework. Instead, a single performance measure, such as θ, is replaced by an estimate of attribute mastery. MYMAP provides an arguably more robust measure of performance that is based on both attribute mastery and IRT ability. The introduction of the *item digraph* illustrates how the interconnectivity of the graph can be exploited for the purpose of optimizing test length. A strength of MYMAP is that it departs from the stochastic item administration relied upon by many, if not all, existing CDAs. By optimizing the stopping conditions, it is found that MYMAP converges more efficiently than RSM-CAT and is comparable in efficiency to POSET-CAT. These results demonstrate how ordering items with a directed graph may answer the call for the sequential item administration much needed for the implementation of an effective CDA.

Evaluation of Content Maps Constructed via a K12 Computer Adaptive Mathematics Assessment

Author
Johnny Denbleyker (Houghton Mifflin Harcourt)

Shuqin Tao (American Institutes for Research)
Ye Ma (University of Iowa)
Mingqin Zhang (University of Iowa)

Abstract
Learning maps (Adjei & Hefferman, 2015) have been designed to detect and target students' knowledge in multiple domains and understand their increasingly sophisticated level of understanding a particular topic. Various methods have been proposed for learning maps. Typically, content experts hand-design learning maps using theory and hypothetical ideas of learning progressions (LPs). However, attempts to validate LPs empirically "have been virtually non-existent" (Kizil, 2015). Validation methods can include a variety of empirical methods (Cen, et al, 2006; Desmarais, et al, 2007). Embedding psychometric properties into the construction of learning maps provide for greater tangible analyses. And, in doing so, could enhance the accuracy in providing individualized learning material to students. The obtained information also may be useful for supporting classroom assessment and targeted instruction for classroom teachers.
Computer adaptive testing (CAT) creates a great opportunity to assist validation efforts for various types of learning maps whether based on theory-driven learning progressions or more quantitatively-driven content maps. In many computer adaptive assessments, typically a grade span of items is involved as part of a large item pool. In item pools so constructed, there exists potentially multiple items to assess each particular content standard. The generalization across content standards, in addition to assessing item performance across grade levels, creates opportunities for a vigorous empirical way to validate how students typically perform on content standards relative to expectations. This is a key benefit CAT has over more traditional fixed-form assessments; CATs allow for greater flexibility with respect to validation efforts.
One specific example of this validation approach via CAT is to utilize a large item bank where examinees are administered the same items nested within defined content standards across multiple grade levels. This research aims to specifically evaluate (1) the Quantile Framework, an established mathematical developmental scale that spans the developmental continuum from Kindergarten mathematics through high school, and (2) the level of invariance on empirical-driven content maps across grade levels.
To address these analytical questions regarding content maps, empirical grade level aggregated item difficulties via item response theory serve as the evaluation component. These empirically derived progressions of content standards, which are formally referred to here as content maps, are compared across separate grade level calibrations and evaluated for consistency and fidelity in both adjacent and nonadjacent grades. A key aspect in content maps and/or learning progressions across grade-levels is that they display a level of invariance across instructional grade levels. The adaptive data administers common instructional standards across grades where items within each standard are empirically calibrated based on examinee grade-levels, which are subsequently aggregated and ordered to form a content map.

Experimentation of a MST for math skills in large scale surveys in Italy.

Author
Emanuela Botta

Abstract
The research is aimed at the construction of a multi-level adaptive test (MST), for the evaluation of the mathematical skills of Italian students of 10th grade, and was carried out in collaboration with Invalsi for a PhD study of "La Sapienza" University of Rome. The research started from the definition of the construct to be measured, taking into account both national and international references. A specific item bank was then built, by carrying out two distinct pre-test phases anchored between them. Sampling of students, distributed throughout Italy and in various types of schools, was always carried out with random assignment of forms to students. In the first phase, 18 test forms (460 items) anchored together were administered to a sample of 4672 students. In the second phase, 11 test forms (403 items) were administered to a sample of 5797 students. These tests were anchored to each other and to those of the previous phase. The selection of the items to be kept in the bank was made by calibrating the difficulty of the items according to the Rasch model, verifying the unidimensionality with specific EFA, and assessing the fit of the bank to the model. The linear scaling of the two banks was carried out, based on the mean and standard deviation of the difficulty parameter of the anchor items, in order to place the values of *b* of items on the same scale. In order to develop an MST 1-3-3 model (Fig.1), three intervals have been identified along the continuity of the ability, with the central interval placed on the mean ability of the sample. For level 1, a routing module (16 items) was built. For levels 2 and 3, two modules were built for each skill interval, consisting of 18 and 12 items.

Multistage Testing Using the D-Scoring Method: Research and Piloting at the NCA in Saudi Arabia

Author
Dimiter M. Dimitrov, Ph.D.
Hanan M. ALGhamdi, Ph.D.
Abdullah A. Alqataee, Ph.D.

Abstract
A recently developed method of test scoring and item analysis, called *D-scoring method* (DSM), is adopted and implemented in automated systems at the National Center for Assessment (NCA) in Saudi Arabia. Under the DSM, the *D*-score of a person is based on his/her response vector of (1/0) scores weighted by the expected difficulties of the test items. The *D*-score (on a scale from 0 to 1) shows what proportion (or %) of the ability required for total success on the test is demonstrated by the examinee. The DSM is a classical analog to IRT in terms of (a) scale intervalness, (b) same scale for persons' scores and item difficulties, and (c) analytic models of item response functions (IRFs), which allow for estimations of true scores, conditional errors of measurement, and other psychometric measures on the *D*-scale.
The use of multistage testing (MST) using the DSM is under current research and piloting at the NCA. The purpose of the proposed presentation is to describe results from MST studies using simulated and real data from the *General Aptitude Test* (GAT) used at the NCA. A recent simulation study (Authors, 2019) examined the efficiency of MST using the DSM under four conditions for test design (fixed linear vs. MST adaptive), module selection, and scoring rules in two scenarios, (a) 1-3 MST structure for a 30-item test (15 items in each stage), and (b) 1-2-3 MST

structure for a 60-item test (with 20 items in each stage). The results suggest that the implementation of MST using the DSM is feasible and can improve the measurement efficiency. All studied conditions with MST outperformed the fixed, linear case in score recovery measured by Pearson correlation between true and observed scores.
Conducted also was a MST study with real data from GAT administrations to high-school graduates in Saudi Arabia The purpose was to examine the suitability of (a) the 2-4-4 MST structure currently used for GAT and (b) the rules for selection of modules and routing of examinees. There were 24 items in each module with the two modules at Stage 1 designed to be equivalent in content and item difficulties. The MST scores were also compared with those from a previous CLT administration GAT for the same examinees on MST-CLT correlations, mean differences, and consistency of criterion-based (pass/fail) classifications of the examinees. The results showed that the transition from CLT to MST administration of GAT is feasible. However, the 2-4-4 MST structure is too complex, with very low frequency of examinees in some paths, and needs to be revised. Problems with the selection of modules and routing of examinees were also signaled and will be discussed with this presentation. The study of MST structures such as 1-2-3 or 1-3-3 for GAT is recommended for the next steps of research on MST applications at the NCA.

A Predicted Error Reduction Stopping Rule for Multidimensional Computer Adaptive Tests

Authors
Scott Morris (Illinois Institute of Technology)
Michael Bass (Northwestern University)
Matthew Lauritsen (Illinois Institute of Technology)
Richard Neapolitan (Northwestern University)

Abstract
Variable-length computer adaptive tests (CATs) allow test length to be tailored to match the characteristics of each examinee. The potential to achieve high measurement precision with a much shorter test is particularly attractive in domains such as patient reported outcomes, where there is a concern about minimizing the burden of measurement on patients. Multidimensional CAT (MCAT) provides additional gains in efficiency by allowing information to be shared across several related traits.
A key component of a CAT is the stopping rule. A common variable-length rule is to administer items until a specified level of measurement precisions is reached (e.g., SE < .3). This approach is effective when all trait levels are well represented in the item bank. However, in some settings, the item bank can be misaligned with the trait distribution, such that there are trait levels at which the available items provide little information. This can be found in assessments of patient reported health outcomes, such as emotional distress (e.g., anxiety, depression) and physical functioning, where items have been primarily developed to differentiate among levels of dysfunction. In such cases, respondents on the positive end of the scale might be asked a large number of questions, without ever approaching the SE cutoff.
Choi, Grady & Dodd (2010) proposed an alternative approach based on predicted SE reduction (PSER). If no item is expected to substantially improve measurement precision, there is no point administering additional items and the exam would stop, regardless of whether the SE cutoff

has been reached. This approach has been found to substantially reduce the number of items administered to individuals for whom the item bank provides little information.
The current research extends the PSER stopping rule to MCAT. MCAT introduces the additional complication that we are attempting to simultaneously maximize precision on multiple traits. We present four versions of the multidimensional error reduction stopping rule, which alternately apply the stopping rule to: a) the trait with the largest SE, sum of SE across all traits, c) the sum of posterior variances across all traits, and the sum of posterior variance across those traits for which the current SE is above a threshold. The relative efficiency and precision of these alternatives were examined using a 3-dimensional CAT for the PROMIS emotional distress banks (depression, anxiety and anger; Cella et al., 2010; Morris Bass & Neapolitan, 2017).
In general, the results support supplementing the minimum SE cutoff with a rule based on predicted error reduction. When there are no items in a bank that are informative for an examinee, the proposed stopping rule will achieve similar precision with substantially fewer items than would be required using the SE cutoff alone. All four methods performed similarly, and all were effective at stopping the MCAT early when additional items were unlikely to improve precision. When differences among methods were found, the Sum of Variance approach performed best, producing comparable precision with slightly fewer items.

Adaptive Multiclassification Testing with Multidimensional Polytomous Items

Authors
Zhuoran Wang (University of Minnesota)
Chun Wang (University of Washington)
David J Weiss (University of Minnesota)

Abstact
The adaptive classification testing (ACT) is a variation of computerized adaptive testing (CAT) that was developed to efficiently classify examinees into multiple groups based on predetermined classification cutoff points. All existing multidimensional ACT studies handled multidimensional classifications in a unidimensional space by performing classification on a composite of multiple traits. However, classification along separate dimensions is sometimes preferred because it provides clearer information regarding a person's relative standing along each dimension. This type of classification is referred to as *grid classification*, as each examinee is classified into one of the grids encircled by cutoff scores (lines/surfaces) on different dimensions. Complications arise when there is more than one cutoff along each dimension. We proposed two new stopping rules for multidimensional ACT with grid classifications, namely, the grid classification generalized likelihood ratio (GGLR) and simplified generalized likelihood ratio (SGLR). We also proposed a new item selection rule, namely, the posterior weighted D-optimal on cutoff points (PWCD-optimal).
A simulation study was conducted to evaluate the performance of multidimensional ACT, using a two-step measurement CAT as a baseline. In the latter scenario, a variable-length multidimensional CAT was conducted, followed by a post-hoc classification. The three-dimensional multidimensional graded response model (MGRM) with four response categories was used. The item bank contained 300 between-item

multidimensional items with 100 items loading on each dimension. Examinees were classified into four groups along each dimension, so there were $4^3 = 64$ classification grids in total. The minimum and maximum test length were fixed at 7 and 60 items. Three item selection methods (the D-optimal, PWCD-optimal, and multidimensional mutual information) and four termination criteria (the GGLR, SGLR, and the between-item multidimensional SPRT and CI) were applied in the grid multiclassification ACT. The D-optimal item selection method and the compound termination criteria (Wang et al., 2018) were used in the two-step measurement CAT. The cutoffs for each stopping rule in the two approaches were selected to carefully yield similar classification accuracy, such that the resulting average test length (ATL) is a useful indicator of test efficiency. Results showed that, when the D-optimal and PWCD-optimal item selection methods were used, ACT resulted in up to 20% shorter ATLs than the two-step approach. In this way, ACT is more efficient than the two-step approach. Among the four termination criteria for ACT, the between-item multidimensional SPRT and CI outperformed the two new stopping criteria.

KEYNOTES:

The Road Ahead: From Computer Adaptive Testing To (Artificially) Intelligent & Holistic Education

Author
Alina A. von Davier (ACTNext by ACT)

Abstract
Computer adaptive tests have paved the way to personalized and adaptive education. Over time, the learning and the testing of learning progress has become more integrated due to advances in technology. Learning and assessment systems (LAS) have grown and taken shape to incorporate concepts from both models for assessment and models for learning. Adaptation and personalization that are based on both machine learning and psychometrics have become the state-of-the-art features of these LAS, transforming them into veritable AI-assistants. Nevertheless, in our work on improving and refining the systems we are adding a new dimension, *navigation*. It important to understand what the capabilities of a learner are and how to grow and expand these capabilities, but we must consider a holistic measurement of the learner as well as where the learner is headed; we need to consider models for navigation. This holistic perspective of learning and assessment systems is encapsulated in the computational psychometrics framework, a framework for designing & developing learning and assessment systems with multimodal data, including navigational components such as behavior and social & emotional learning features. Computational psychometrics blends theory-driven psychometrics with machine learning and facilitates the abstraction from rich, raw data to conceptual models. The new version of AI-assistants is getting closer to an interactive coach that can assist the learner and the teacher in supporting the learning goals. Several examples of research projects are provided.

Recent advances in cognitive diagnosis computerized adaptive testing

Author
Jimmy de la Torre (University of Hong Kong)

Abstract
In recent years, cognitive diagnosis models (CDMs) have gained increasing popularity because of their potential to provide finer-grained inferences that can inform learning and teaching. To further capitalize on the advantages of CDMs and make diagnostic testing more efficient, cognitive diagnosis computerized adaptive testing (CD-CAT) has been proposed. This presentation will compare various item selection indices that have been used in CD-CAT with respect to attribute classification accuracy, item usage, and implementation time. To examine the benefits of CD-CAT with real data, its attribute classification accuracy vis-à-vis a proportional reasoning test will be compared with those of the original and optimally designed paper-and-pencil versions of the test. The presentation will also discuss other recent advances in the area, which include item selection indices for different item responses and test formats, nonparametric CD-CAT, and CD-CAT implementation when the number of attributes is large.

The CAT Curmudgeon: Some Thoughts from 50 Years of CAT

Author
David Weiss (University of Minnesota)

Abstract
CAT was originally operationalized in the early 1970s and is the first application of artificial intelligence in psychological and educational testing. Over 50 years, CAT has matured in a number of ways and is now advancing measurement throughout the world. Yet in a number of respects, implementation of CAT and the IRT that support it are still tethered to old ways of doing things, thus limiting the capability of CAT to optimally move measurement forward. I will identify a number of these issues and suggest ways of moving forward so that CAT can unleash its full potential to improve measurement.

Multidimensional Computerized Adaptive Testing in Health Measurement: Lessons Learned

Author
Chun Wang (University of Washington)

Abstract
There is an increasing interest in integration of patient's perspectives in clinical research, as indicated by the growing emphasis in developing patient reported outcome (PRO) measures. Computerized adaptive testing (CAT) has become a useful tool to deliver PRO measures. Moreover, because many of the PRO measures are correlated, quite a few studies have shown that, unsurprisingly, multidimensional computerized adaptive testing (MD-CAT) improves measurement efficiency compared to separate unidimensional CATs. In this talk, I will share an ongoing 5-year collaborative project on developing a MD-CAT for measuring post-acute care patients' rehabilitative care needs. New challenges emerge, including how to stop a MD-CAT at the right time when the item bank may be exhaustive of appropriate items? How to design an efficient stopping rule when the purpose is to classify individuals into multiple categories? Viable solutions are proposed and they have been successfully implemented in a live MD-CAT. It is expected that the lessons we learned and the solutions we provided offer insights into the future design of MD-CATs.

Computerized Adaptive Testing with Response Revision: Challenges, Solutions and Applications

Author
Shiyu Wang (University of Georgia)

Abstract
Test takers' response revision behavior can be commonly observed in a paper-pencil or computer based linear test. In general, it's believed that response revisions are able to correct errors made by careless or speeded responding. Many studies support response revision when there is a good reason for doing so. Research on response revision in computer-based adaptive tests is limited. This is mainly because many adaptive tests do not allow test takers to review and revise responses during the test. In this work, the challenges of allowing response review and revision in computerized adaptive testing (CAT) are discussed first. The major issue is on maintaining a robust and efficient CAT system while allowing response revision. In fact, if response revision is allowed in CAT, until a response is changed it is used for the interim ability estimation. As a result, the revision influences the item selection process. This causes concerns that allowing response revision in CAT may encourage test-taking strategies intended to exploit the adaptive item selection process, which tends to administer easier items as the estimated ability decreases, thereby causing biased ability estimates.

Expanding the Meaning of Adaptive Testing to Enhance Validity

Author

Steven L. Wise (NWEA)

Abstract
Traditionally, computerized adaptive tests (CATs) provide efficient testing by administering items whose difficulties are matched to test taker ability. That is, the primary focus is on score precision. I will propose that it is time to expand the ways our tests adapt by re-focusing our attention on score validity. Specifically, I believe that future CATs will increasingly incorporate methods and features that can control the impact of construct-irrelevant factors that diminish validity, such as disengagement, test anxiety, and cheating. To demonstrate this idea, I will discuss the test-taking features that have been added to NWEA's MAP Growth assessment.