2024 IACAT Conference Program Book

IACAT

9th Conference of
the International Association
for Computerized Adaptive Testing
September 24-27, 2024

연세대학교
YONSEI UNIVERSITY

Contents

Welcome Message from the IACAT President-Elect Wim J. van der Linden

Dear attendants,

It is with great pleasure than we welcome you to the 2024 Conference of the International Association for Computerized Adaptive Testing, which will take place on the hospitable campus of Yonsei University, Seoul, South Korea.

The main theme of the conference is Improving Assessment with Adaptivity and Artificial Intelligence. With seven keynote addresses, over ninety individual paper presentations, three symposia, three workshops, a welcome reception, banquet, and a choice from several social activities, we are already well on our way to honor this theme.

We are extremely thankful to Professor Ji Hoon Ryoo, our host during the conference, along with his local organizing committee, who have done their utmost to guarantee a meeting that promises to be exciting and instructive.

But the success of the meeting will also depend on you as a participant. So please, do participate in our scientific discussions, reconnect with colleagues you've met before, and feel to share a drink or a meal with someone you don't know yet. This will be most rewarding, not only to yourself but to the other participants as well.

We look already forward to seeing you in Seoul.

Wim J. van der Linden

Incoming President

Welcome Message from the Local Organizing Chair
Ji Hoon Ryoo

Dear Colleagues and Conference Delegates,

I am delighted to welcome you to the 2024 IACAT Conference in Seoul, South Korea. It is a great honor to host the world's leading researchers in Computerized Adaptive Testing (CAT) at Yonsei University. Established in 1885 with Korea's first modern hospital, "Chejungwon," and the missionary school, "Yonhi College," Yonsei University has a rich history. Over the past 70 years, the Department of Education has trained educators, administrators, and researchers, with a focus on educational measurement, psychological assessment, and psychometrics. On behalf of our department, the university, and South Korea, I extend our warmest welcome to you.

In recent years, the field of CAT has broadened, both in scope and quality, particularly with the rise of adaptive learning. The 2024 IACAT Conference will showcase the most comprehensive and cutting-edge research in CAT, with many proposals demonstrating the latest advancements. We are also excited to see young scholars contribute their vision for the future of CAT over the next two years. In addition to presentations based on traditional IRT-based methods for measuring individual differences, this year's program features innovative work that bridges the gap between CAT, artificial intelligence, and adaptive systems that integrate learning and testing.

The conference will kick off on Tuesday, September 24, with three parallel workshops. Following the workshops, we invite you to join us for a Welcome Reception at the Yonsei University Alumni Association Building from 6:00 p.m. to 9:00 p.m. This reception is complimentary for all registered attendees. The first full day of the conference, September 25, will culminate in a Conference Banquet and Presidential Address from 5:10 p.m. to 9:00 p.m. We are honored to have five distinguished keynote speakers, as well as a presentation for the Early Career Award. We wish you a conference with interesting academic exchanges, stimulating social interactions and nice impressions of Seoul, South Korea.

The entire organizing committee is excited to have you with us. We hope this conference offers you enriching academic discussions, stimulating social interactions, and lasting impressions of Seoul, South Korea.

Ji Hoon Ryoo

(Conference Host)

Organizing Committees

Local Organizing Committee

Ji Hoon Ryoo (Chair) / Yonsei University, South Korea

Guemin Lee / Yonsei University, South Korea

Hyun-Jeong Park / Seoul National University, South Korea

Hyo Jeong Shin / Sogang University, South Korea

Hyesung Shin / Korea Institute for Curriculum and Evaluation (KICE), South Korea

Organizing Committee / IACAT Board of Directors

Tony Zara (President) / Pearson VUE

Wim J. van der Linden (President - Elect) / University of Twente

Duanli Yan (Secretary) / Educational Testing Service (ETS)

Nathan Thompson (Membership Director) / Assessment Systems Corporation (ASC)

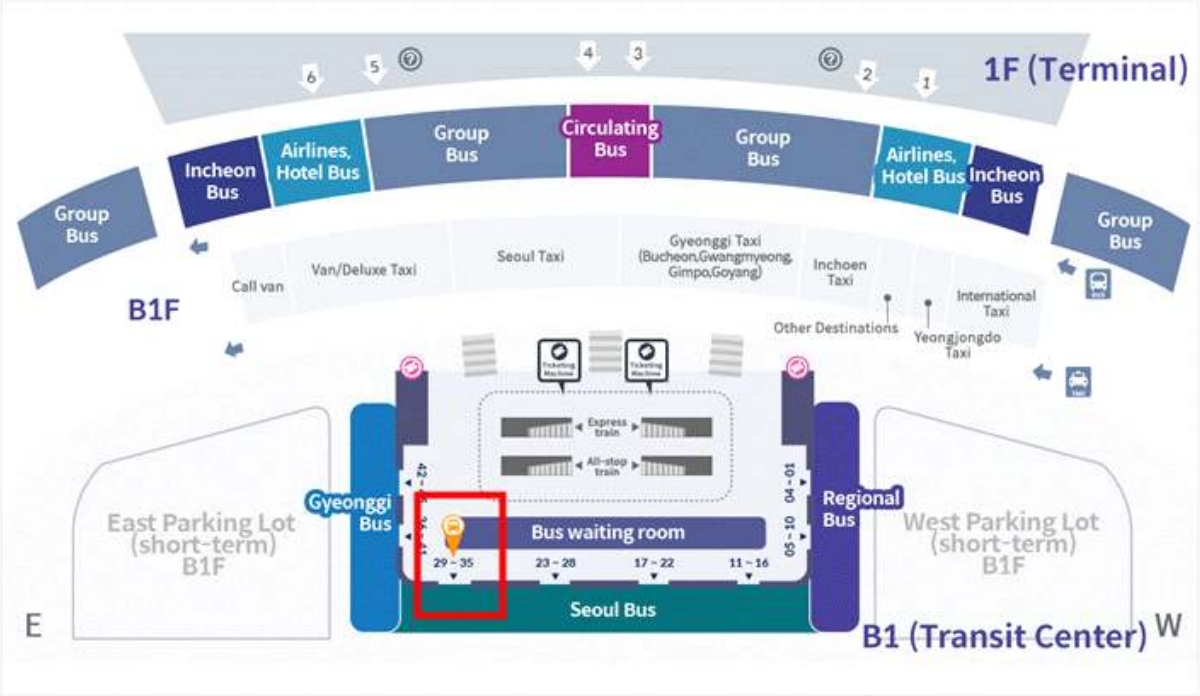
For Your Orientation

Public Transportation Information (From Incheon International Airport)

1. By Bus (6011 Bus / Incheon International Airport to Yonsei Univ.)

Terminal 1 Bus stop location (1st Floor No.5)

Terminal 2 Bus stop location (Transportation center B1 No.31)

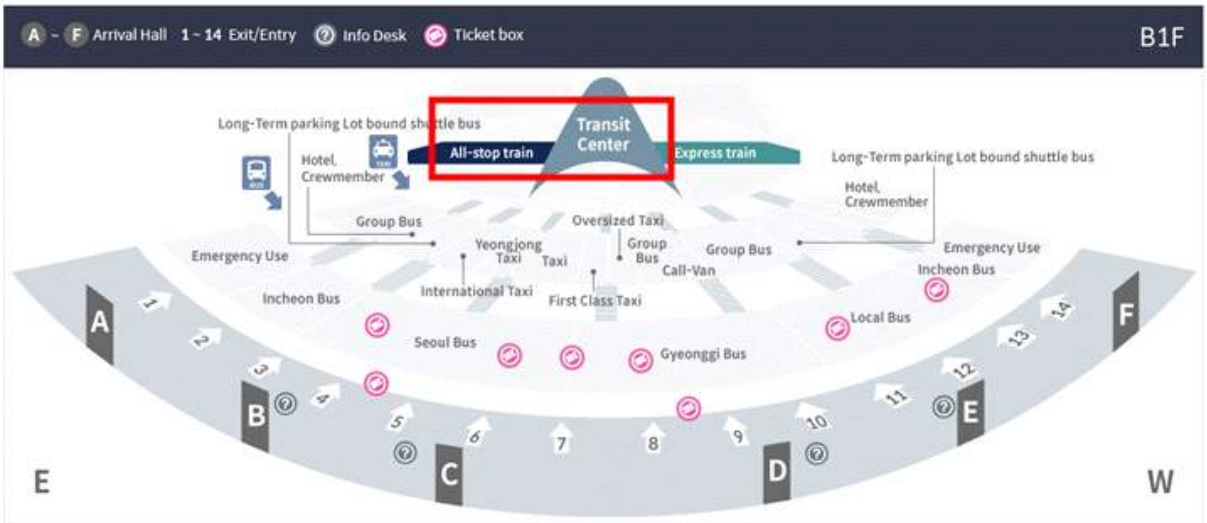


2. By Airport Railroad & Subway

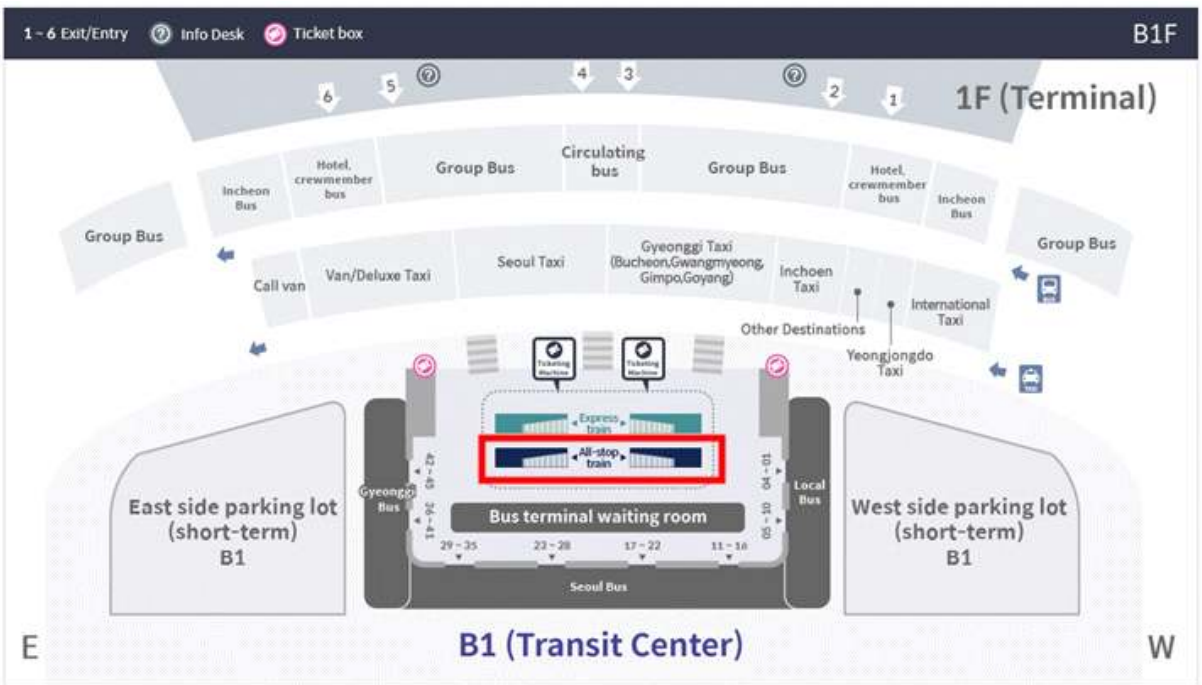
Take Airport Railroad (All-stop train) and transfer to subway line 2 at Hongik Univ. station. Travel one stop and take off at Sinchon Station.

From Sinchon Station Exit 2, walk to Baekyang-ro, Yonsei University. It takes about 10 ~ 15 minutes.

Terminal 1 Airport Railroad location



Terminal 2 Airport Railroad location

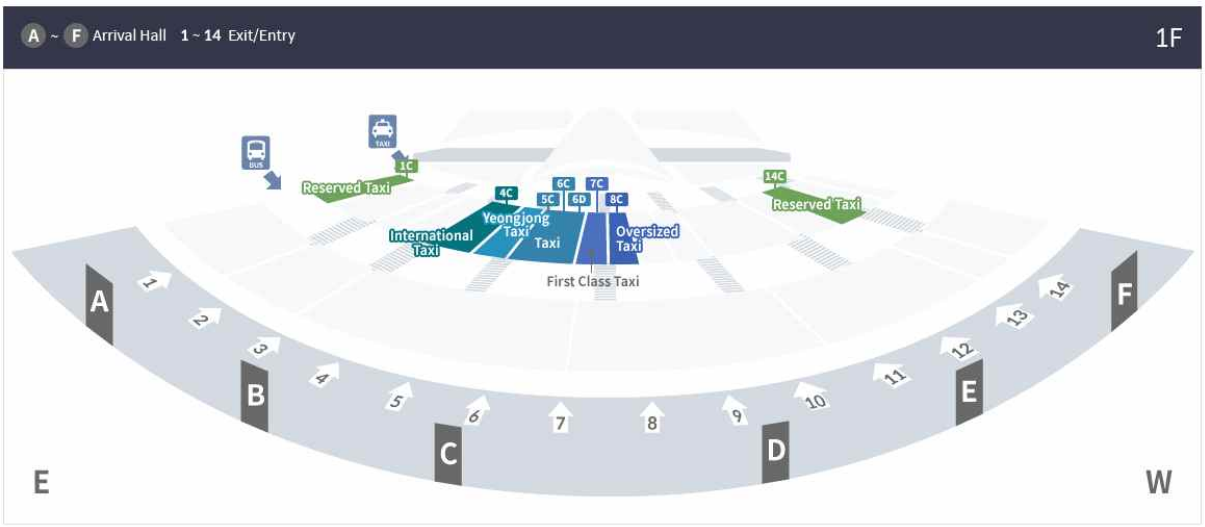


3. By Taxi

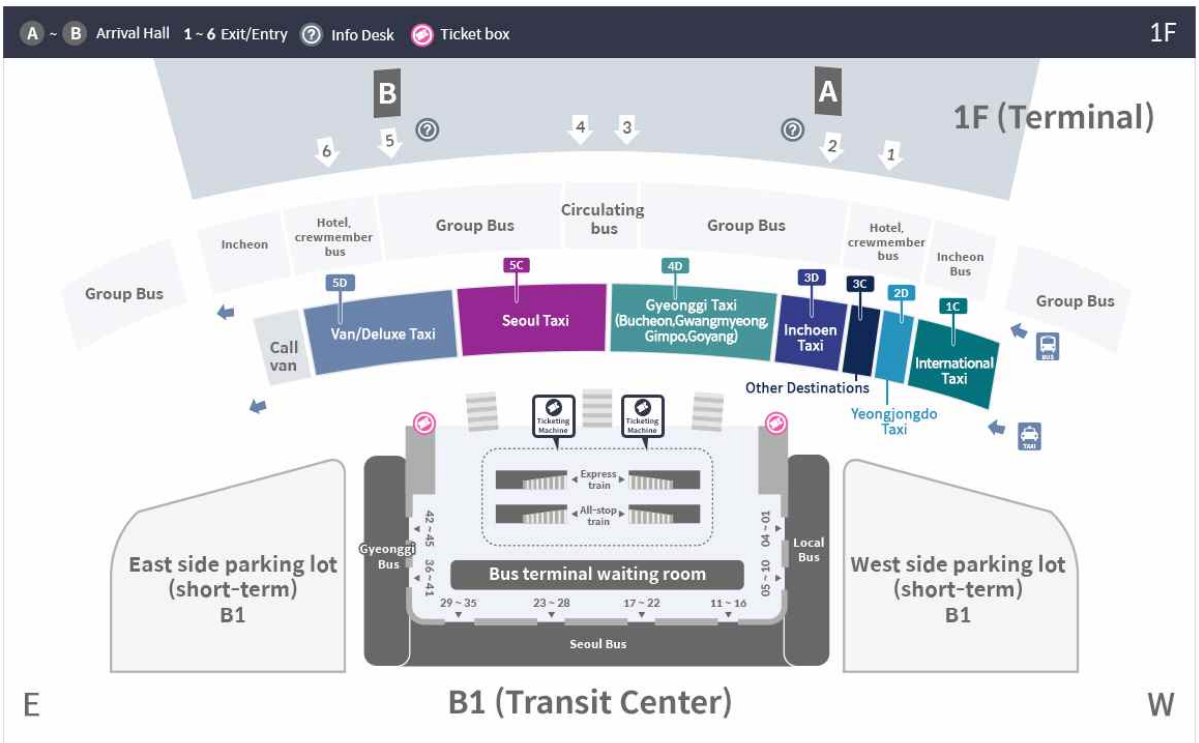
Regular Taxi (Seoul)

International Taxi (https://www.intltaxi.co.kr/ Tel. 1644-2255)

Terminal 1 Taxi stop location

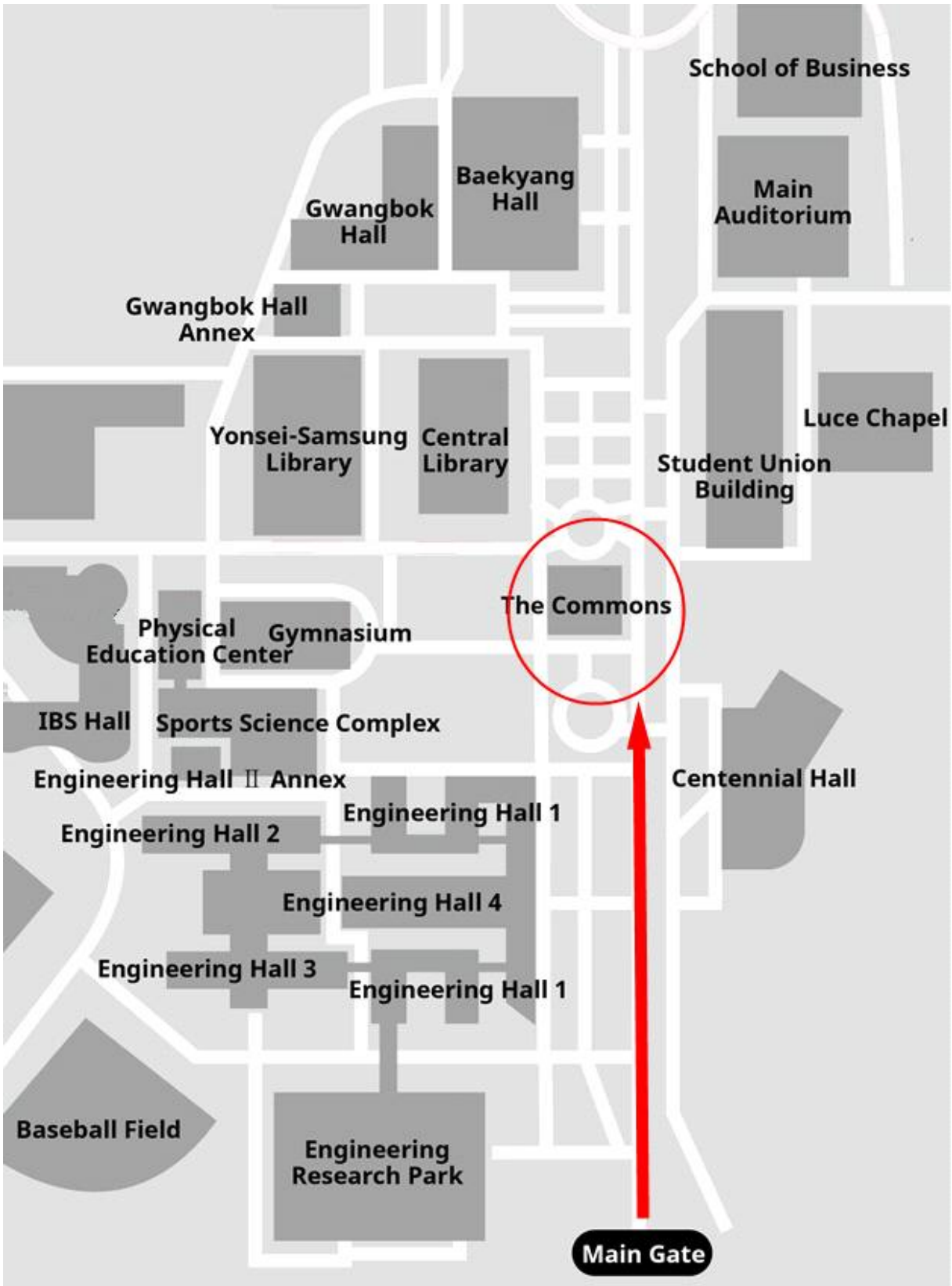


Terminal 2 Taxi stop location



Conference Venue - Yonsei University

Conference Venue The Commons at Yonsei University

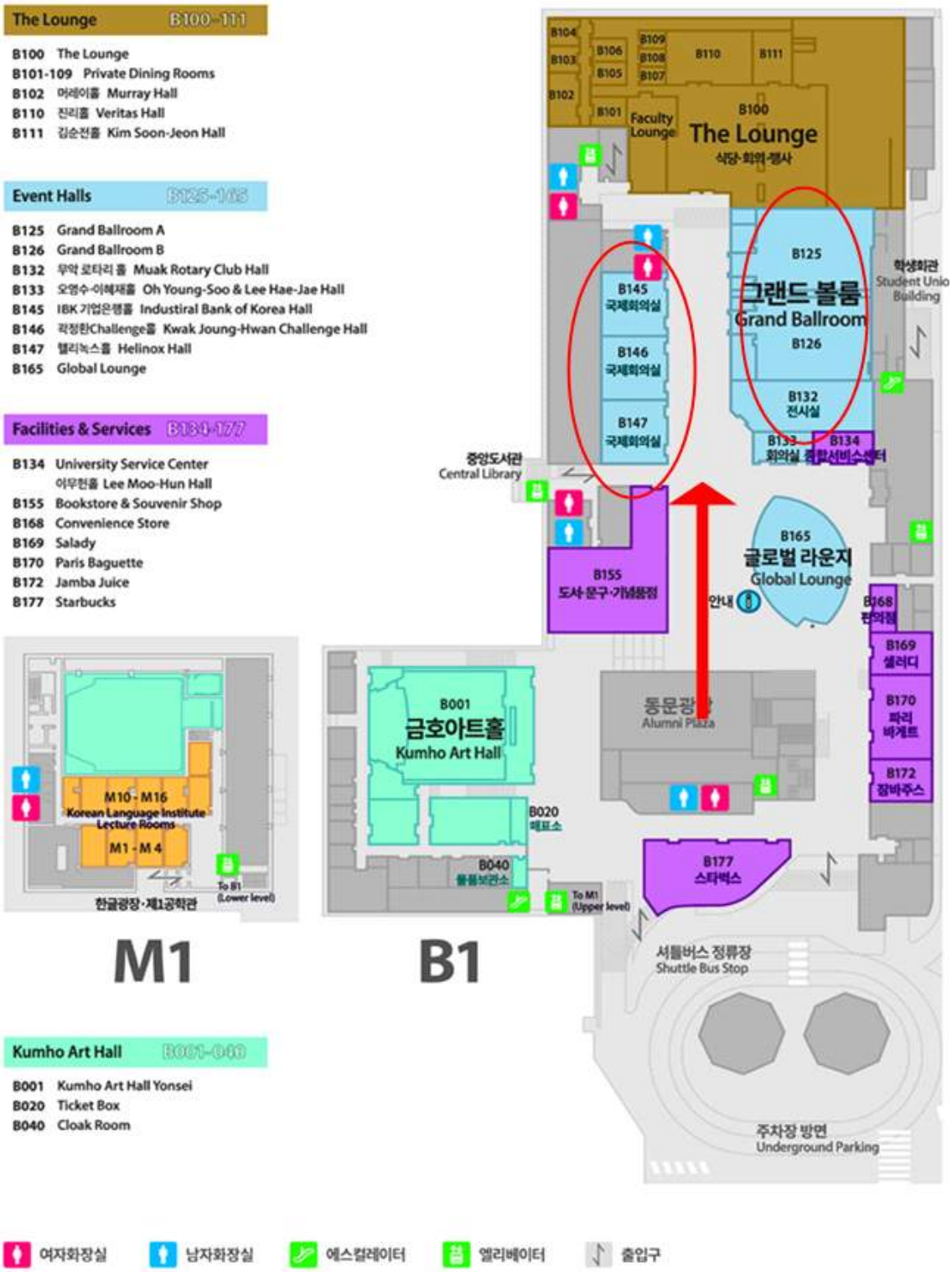


Room Map

Main Lecture : B125 and B126 (Grand Ballroom)

Sessions : B145, B146, B147 (International Meeting Rooms~~Conference Hall~~)

Refreshment : B132 (Rest Area)



Events

Welcome Reception

On Tuesday, September 24, 6:00 PM - 9:00 PM we look forward to welcoming you to a reception with Buffet-style dinner in the Yonsei University Alumni Association Building. This reception is complimentary for all registered attendees.

Banquet

The conference dinner will take place at the Grand Ballroom on Wednesday, September 25, 6:30 PM - 9:00 PM. It is with great pleasure that we invite you to the Banquet in celebration of the successful completion of 9th IACAT conference. This event is an opportunity for all attendees to come together, celebrate academic achievements, and foster further intellectual exchange. Tickets need to be purchased for the banquet using the registration system.

Social Events – Thursday, September 26, 2:30 PM ~ 8:30 PM

Social Event #1:

N Seoul Tower, Myeong-dong, Itaewon ($100 / Dinner is not included!) IACAT 2024

Join us for an unforgettable tour of some of Seoul's most iconic attractions! Our first social event will take you to the heights of N Seoul Tower, through the vibrant streets of Myeong-dong, and into the multicultural hub of Itaewon. This is your chance to immerse yourself in the rich culture and lively atmosphere of Seoul while making lasting memories with fellow participants.

Event Schedule

Time	Activity	Description
2:30 PM	Meet @ Yonsei's Main Gate	Start!
3:00 PM - 5:00 PM	N Seoul Tower Cable Car	Experience the breathtaking views from the top of N Seoul Tower as you ascend by cable car.
5:00 PM - 6:00 PM	Myeong-dong	Explore the bustling streets of Myeong-dong, famous for its street food and shopping.
6:30 PM - 7:30 PM	Itaewon	Visit the multicultural district of Itaewon, known for its diverse range of shops and restaurants.
8:30 PM	Return to Yonsei University	Head back to Yonsei and relax after a day full of exciting experiences.

Social Event #2:

Gyeongbokgung Palace, Hanbok Experience, Seoul Sky ($100 / Dinner is not included!) IACAT 2024

Join us for an unforgettable tour of some of Seoul's most iconic attractions! Our second social event will take you through the historic Gyeongbokgung Palace where you'll get to experience traditional Hanbok attire, up to the heights of Seoul Sky for breathtaking panoramic views, and back to Yonsei University. This is your chance to immerse yourself in the rich culture and lively atmosphere of Seoul while making lasting memories with fellow participants.

Event Schedule

Time	Activity	Description
2:30 PM	Meet @ Yonsei's Main Gate	Start!
3:00 PM - 5:30 PM	Gyeongbokgung Palace & Hanbok Experience	Explore the majestic Gyeongbokgung Palace and experience the traditional Korean attire, Hanbok.
6:30 PM - 7:30 PM	Seoul Sky	Take in the stunning panoramic views from the observation deck of Seoul Sky..
8:30 PM	Return to Yonsei University	Head back to Yonsei and relax after a day full of exciting experiences.

Schedule at a Glance

The conference consists of workshops, keynote lectures, invited symposia, a submitted symposium and thematic paper sessions. Lunch, snacks, cold and hot drinks during the conference and the welcome reception on day 1 are covered by your registration fee.

The Schedule at a glance on the next four pages gives an overview over the IACAT 2024 conference. The sessions are numbered consecutively. The abstracts for the individual presentations within these sessions can be found in the pages following the Schedule at a Glance. The abstracts are ordered by the session numbers.

Day 0

	Session 0 (Day 0, 1:00 - 5:00)			
Session0	S0-1 (B145)	S0-2 (B146)	S0-3 (B147)	Grandball room
	Workshop 1: Introduction to AI-based Automated Item Generation and Scoring in Adaptive Testing	Workshop 2: Introduction to IRT and CAT	Workshop 3: The Shadow-Test Approach to Adaptive Testing	
1	Duanli Yan (ETS)	Nathan Thompson (ASC)	Seung W. Choi (UT-Austin)	
2	Alina A. von Davier (Duoling)		Wim J. van der Linden (Univ. of Twente)	
Keynote: Ji Hoon Ryoo (Day 0, 6:00 - 6:30) & Welcome Reception (6:30 - 9:00)				

Day 0

Opening Ceremony (9:00 - 9:30) & Keynote: Peter W. van Rijn (Day 1, 9:30 - 10:30)

Day 1

	Session 1 (10:40 - 12:00)			
Session1	S1-1 (B145)	S1-2 (B146)	S1-3 (B147)	S1-4 / Symposium 1 (Grandball room)
	AIG, Automative Scoring 1	CAT Applications 1	Advancements in Adaptive Testing	Advancements in Bayesian Methods for CAT
1	Syed Abdul Hadi	Yeonwho Kim	Alina Von Davier	Andreas Frey
2	Jyoti Pandey	Dong Gi Seo (Dogyeong Kim)	Angela J. Verschoor	Luping Niu (Seung W. Choi)
3	Joe Watson	Jungsoo Kim	Vipin K Chilana	Ae Kyong Jung
4	Mayank Kumar		Kristof Kovacs	Jonathan Templin

	Session 2 (2:00 - 3:20)			
Session2	S2-1 (B145) / Symposium 2	S2-2 (B146)	S2-3 (B147)	Grandball room
	Research for Practical Issues and Solutions in Computerized MST	Item Selection Methods 1	Cognitive Diagnosis CAT 1	
1	Nathan Thompson	James Sharpnack	Junsik Sim	
2	David Weiss (Duanli Yan)	Haejin Kim	He Peng	
3	Mark Reckase	César Antonio Chávez Alvarez (Nathan Thompson)		
4	Kyung T. Han			
5	Victoria Song			
6	Alina A. von Davier			

	Session 3 (3:40 - 5:00)			
Session3	S3-1 (B145)	S3-4 / Symposium 6 (B146)	S3-3 (B147)	S3-2 / Symposium 4 (Grandball room)
	AIG, Automative Scoring 2	Remembering Theo Eggen: A Symposium to Honor His Intelletual Legacy	AI Topics 1	Integrating AI into Adaptive Testing
1	Bartosz Kondratek	Bernard Veldkamp	Eun Hye Ham	Hyun Suk Ryoo
2	Jinmin Chung	Nathan Thomson	Mikyoung Yim	Yeongjin Jo
3	Mayank Kumar	Maaike van Groen	Xiangen Hu	Hyo Jeong Shin
4		Angela Verschoor		Seewoo Li

Presidential Address: Wim. van der Linden (Day 1, 5:10 - 6:00) & Banquet (6:30 - 9:00)

Day 1

		Session 4 (8:30 - 9:50)			
Day 2	Session4	S4-1 (B145)	S4-2 (B146)	S4-3 (B147)	S4-0 / Symposium 3 (Grand Ballroom/Hybrid)
		Item Banking	CAT Applications 3	AI Topics 2	Current CAT Research at the U of MN
	1	Ren Jie	Dorinde Korteling	Young Koung Kim	Raj Wahlquist
	2	Yuan Ge	Nathan Thompson	Ryan Lerch	Matthew A. Snodgress
	3	Rae Yeong Kim	Istiani, M.Psi	Mfonobong Umobong	Ming Him Tai
	4	Haejin Kim			Joseph N. DeWeese
	5				Robert Chapman
	6				Jesus Delgado
		Keynote: Chia-Ling Hsu (Day 2, 10:00 - 11:00)			
		Session 5 (11:10 - 12:10)			
	Session5	S5-1 (B145)	S5-2 (B146)	S5-3 (B147)	S5-4 / Symposium 5 (Grandball room)
		IRT Applications	Cognitive Diagnosis CAT 2	Constraint Management in CAT	The Development of iSKA
	1	Ryan EK Man	Ivy P. Mejia	Kylie Gorney	Yongsang Lee
	2	Aikorkem Zhapparova	Ahoo ShokraieFard	Chia-Wen Chen	Hwanggyu Lim
	3				Kyung T. Han
	4				Dongkwang Shin
		Keynote: Seung W. Choi (Day 2, 1:30 - 2:30)			

Day 3

	Session 6 (8:30 - 9:50)			
Session6	S6-1 (B145)	S6-2 (B146)	S6-3 (B147)	S6-4 (Grandball room)
	CAT Applications 4	Multi-stage Testing 1	Software & Systems related to Adaptive Testing	Machine Learning
1	Pradyumna Amatya	JP Kim	ChaeEun Kim	Nathan Thompson
2	Selma ŞENEL	Insub Shin	Mayank Kumar	Lihua Yao
3	Mehmet Can Demir	Garrett Ziegler		Young Jin Kim
4	Haniza Yon	Kyoungwon Bishop		

Keynote: Marlana Curi (Day 3, 10:00 - 11:00)

	Session 7 (11:10 - 12:10)			
Session7	S7-1 (B145)	S7-2 (B146)	S7-3 (B147)	Grandball room
		Item Selection Methods 2	CAT in Reality	
1		Rodrigo S. Kreitchmann	Arvind Singh	
2		Jinha Kim (Dong Gi Seo)	Xiaowen Liu	
3		Um i Lela	David Budzyński	

Keynote: Kyung T. Han (Day 3, 1:30 - 2:30)

	Session 8 (2:40 - 3:40)			
Session8	S8-1 (B145)	S8-2 (B146)	S8-3 (B147)	S8-4 (Grandball room)
	Multi-stage Testing 2	AI Topics 3	Other Adaptive Testing Topics	CAT Applications 2
1	Hanan AlGhamdi	Artur Pokropek	Semih Topuz	Jeongwook Choi
2	Jinmin Chung	Burhanettin Ozdemir	Luz Bay	Minjung Kim
3		Kim Seong Il	Yoojin Chelsea Jang	Jeongin Cha

Closing Ceremony (Day 3, 4:00 - 4:30)

Day 3

Abstracts

S0-1: Workshop 1 - Introduction to AI-based Automated Item Generation and Scoring in Adaptive Testing

Duanli Yan & Alina A. von Davier

Abstract: This workshop introduces a novel framework, "the item factory", for managing large-scale test development including automation of item generation, quality review, quality assurance, and crowdsourcing techniques in adaptive testing. We will present an overview of the latest natural language processing (NLP) techniques and large language models for automatic item generation, alongside evidence-centered design and psychometric principles and practices for test development. We will discuss the application of engineering principles in designing efficient item production processes (Luecht, 2008; Dede et al, 2018; von Davier, 2017).

S0-2: Workshop 2 - Introduction to IRT and CAT

Nathan Thompson

Abstract: This workshop provides a broad overview of item response theory (IRT) and computerized adaptive testing (CAT) for those who are newer to the field. We assume a knowledge of basic psychometrics such as classical test theory. The workshop begins with a background on (IRT), how it can be used to evaluate item and test performance, and how it provides a number of improvements over the classical approach. We then provide an introduction to CAT, describing the components and algorithms necessary to build an effective CAT program: item bank calibrated with IRT, starting point, item selection rule, scoring method, and termination criterion. Finally, we discuss important aspects regarding how you might evaluate and implement CAT for your organization.

S0-3: Workshop 3 - The Shadow-Test Approach to Adaptive Testing

Seung W. Choi, & Wim van der Linden

Abstract: This short course has four different sections. In the first section, we explain the ideas underlying the shadow-test approach, discuss a few practical aspects of its implementation, and show some of its generalizations to different test formats. The next two sections present two applications of the approach illustrating its versatility. One application is adaptive testing with a mixture of discrete items and items organized as sets around a common stimulus with continued adaptation within each set. The other is adaptive testing with field-test items inserted for adaptive calibration. The final section of the course demonstrates software available for the implementation of the shadow-test approach to adaptive testing and offers the participants hands-on experience with the R package TestDesign. In addition, a brief introduction will be given to Optimal CAT, a currently freely downloadable microservice available for easy integration with current test delivery systems.

S1-1: Paper Session- AIG, Automative Scoring 1

Chair: *Syed Abdul Hadi*

Syed Abdul Hadi

Automated Essay Scoring: Syntactic and Semantic Feature Engineering for Interpretable and Content Aware Systems

Abstract:

Objective: This research explores an automated essay scoring (AES) system for WIDA's K-12 English language assessments. Our novel hybrid approach augments an interpretable syntactic feature set with a content-aware Semantic Feature Set (SFS) leveraging transformers to measure the coherence of student responses with the prompt. By integrating SFS, we aim to build a generalizable AES system that is interpretable, requires minimal data, and adequately penalizes off-topic responses. For piloting, we use grade 9-12 student responses in the WIDA ACCESS writing test comprising short and medium-length responses (150-500 words).

Background: AES evaluates student responses to specific prompts, offering a cost-effective and time-effective solution for grading large volumes of essays and providing consistent and immediate feedback. Transformer-based AES systems, relying on pre-trained language models, have demonstrated considerable success in evaluating long-text and on-topic responses, as shown by Yang et al. (2020) and Wang et al. (2022). However, these systems lack interpretability and fail to penalize off-topic responses adequately. Kabra et al. (2022) highlighted over-stability in many AES algorithms, with models not penalizing or even increasing scores for off-topic responses. Parekh et al. (2020) demonstrated that state-of-art AES systems suffer from adversarial inputs because pre-trained language models are not semantically aligned with real-world knowledge. In their systematic literature review of AES systems, Ramesh et al. (2021) highlight the dearth of research on content-based evaluation. Do et al. (2023) employs a prompt-aware transformer-based approach and achieves state-of-the-art results but cannot address interpretability issues.
This research addresses the need for an AES system that is interpretable, prompt-aware and requires minimal on-topic hand-scored data given its often limited availability.

Methodology: Our hybrid approach augments syntactic features with a Semantic Feature Set (SFS) in a regression-based system to assign a holistic score while ensuring interpretability of individual features. The WIDA scoring rubric assigns a combined holistic score to gauge characteristics like vocabulary, grammar, complexity, discourse, task-awareness, and prompt coherence. Our model extracts 40 syntactic features to gauge grammar, vocabulary, complexity, and discourse. This feature set is then augmented with SFS to measure task-awareness and prompt-coherence. SFS comprises 3 distinct features:

1. Cosine distance of prompt embedding with essay embedding.
2. Aggregated Cosine distance of prompt embedding with sentence-level embeddings.

3. Aggregated Cosine distance of consecutive sentence embeddings using a sliding window of 3 sentences.

To counter noise from lengthier prompts, we employ the BART ConditionalGenerator by Lewis et al. (2020) to condense the prompt. BART utilizes a self-attention mechanism to emphasize key aspects in text. For generating embeddings, we tested BERT, AlBERT, ClinicalBERT, and RoBERT.

Preliminary Results: We use Quadratic Weighted Kappa (QWK) to evaluate the AES system. Preliminary results using just the 40 syntactic features exhibit high agreement between machine and human scores. Initial testing of the syntactic feature model strictly using on-topic responses from the publicly available ASAP dataset attained 94.4% of the human-human rater agreement despite redacted content in the data. Preliminary testing on WIDA's unredacted writing data with the augmented feature set is expected to show significant improvement in performance, specifically on off-topic responses.

References

Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. 2020. Enhancing Automated Essay Scoring Performance via Fine-tuning Pre-trained Language Models with Combination of Regression and Ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1560–1569, Online. Association for Computational Linguistics.

Yongjie Wang, Chuang Wang, Ruobing Li, and Hui Lin. 2022. On the Use of Bert for Automated Essay Scoring: Joint Learning of Multi-Scale Essay Representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3416–3425, Seattle, United States. Association for Computational Linguistics.

Anubha Kabra, Mehar Bhatia, Yaman Kumar Singla, Junyi Jessy Li, and Rajiv Ratn Shah. 2022. Evaluation Toolkit For Robustness Testing Of Automatic Essay Scoring Systems. In Proceedings of the 5th Joint International Conference on Data Science & Management of Data (9th ACM IKDD CODS and 27th COMAD) (CODS-COMAD '22). Association for Computing Machinery, New York, NY, USA, 90–99. https://doi.org/10.1145/3493700.3493765

Heejin Do, Yunsu Kim, and Gary Geunbae Lee. 2023. Prompt- and Trait Relation-aware Cross-prompt Essay Trait Scoring. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1538–1551, Toronto, Canada. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Ramesh, D., and Sanampudi, S.K. 2022. An automated essay scoring systems: a systematic literature review. Artificial Intelligence Review, 55, 2495-2527.
https://doi.org/10.1007/s10462-021-10068-2

Jyoti Pandey

Hybrid Neural Network and Feature Engineering Model for Enhanced Automated Essay Scoring

Abstract: This research explores an automated essay scoring (AES) system for WIDA's K-12 English In computer adaptive testing, scoring multiple-choice items is straightforward. However, incorporating descriptive or short-answer questions requires advanced techniques for automated text scoring. We have seen a significant evolution of Automated Essay Scoring (AES) over the past few decades, from initial rule-based systems to more advanced machine learning and now to sophisticated neural network models. This paper explains the journey of the Institute of Banking Personnel Selection (IBPS), a large-scale testing organization, through these phases and experimentation with a new hybrid model that combines feature engineering with neural networks.

In 2021, IBPS implemented a rule-based model for scoring descriptive papers in some of the large-scale examinations (more than 60,000 essays). This model was found to be rigid and restricted in comprehending the complexity of human language, despite being interpretable. We also discovered that applicants occasionally attempt to manipulate the scoring system in order to gain higher marks. Through our analysis, we have identified several patterns in these attempts.

It was crucial to address these limitations; hence, in 2023, IBPS developed a machine learning model. This model was an improvement over the rule-based model and offered improved flexibility and accuracy. Although this model has its own limitations, it required extensive labeled data and training for each new essay prompt. This model needs good-quality largescale training data to perform better, and its ability to score accurately depends on that, but there is always a trade-off between quality training data and large training data, which poses a challenge for the machine learning model and forces us to look for other options.

Another model we tried was a neural network model. Its main advantage was that this model does not need labeled data. Although this model showed some encouraging advancements in capturing the nuances of natural language and was also able to overcome some limitations of machine learning models, generalizing this model for each prompt was still a problem. It was better at accuracy than a rule-based model, but some of the prompt-specific features were still missed by this model.

To make our model more sensitive to language intricacy and still able to identify prompt specific features, we adopted a new approach in which we utilized both feature engineering

and neural networks. This hybrid model incorporates features based on language, correctness, style, and topicality into neural network models. Techniques such as LDA, TFIDF, cosine similarity, and Networkx were utilized to better understand and evaluate the text content. The objective was to identify the optimal combination of features and models. Though there are some hybrid models available, this model uses some new combinations for feature engineering and was tested on a large amount of real text data. The initial phase of testing of this model indicates that this hybrid model not only improves the scoring accuracy of the text but also enhances the model's

ability to generalize across different topics and writing styles, which was the aim of this model.

This paper provides a detailed account of the model development process, including the
challenges encountered, the solutions implemented, and the results achieved. It contributes to the field of AES by offering a novel framework that balances the interpretability of rule-based systems with the flexibility and accuracy of machine learning and neural network approaches. The findings suggest that the integration of feature engineering with advanced neural network architectures holds significant promise for the future of automated essay scoring.

Joe Watson, Chia-Wen Chen, Ivan O'Conner, & Luning Sun

Automated Essay Scoring for Measuring Psychological Constructs

Abstract: Automated Essay Scoring (AES) methods commonly involve applying construct-specific rubrics developed by domain experts or supervised machine learning (ML) models trained to predict human ratings. However, for many constructs, obtaining relevant rubrics or human labelled essays is challenging. Our research introduces innovative AES methods to support the estimation of psychological traits, which do not require specific rubrics or essay marks from human raters. These methods include: using supervised ML to predict total scores for a subset of items based on essay features; applying a generic rubric, where response values are derived from a document term matrix; and, utilizing generative artificial intelligence (AI) to assign essay scores across multiple criteria. Following the generation of responses through an AES technique, an item response theory model is applied to the combined responses from closed items and those obtained through AES. AES items with low discrimination or local dependence are iteratively removed until a final model is established.

Our study focuses on measuring locus of control (LoC) among US university students. AES approaches are developed using synthetically generated responses (n=1,000, theta mean=0,theta sd=1) to 29 binary items, reflecting an established LoC test, and an essay item on the extent to which the respondent considers their current life position to be a product of fate. Responses to closed items are simulated for a set of artificial item parameters. Essays are created using generative AI, with model prompts varying according to each respondent's theta and randomly assigned characteristics (age, degree topic, hobby, and US state of origin).

We assess whether our novel AES methods can enhance test performance in an adaptive testing context. Adaptive tests including responses from closed items and various approaches are compared against each other and a baseline adaptive test (without A response), using a separate set of synthetic responses to the 29 closed items and essay item(n=1,000, theta mean=0, theta sd=1). Our findings indicate that several self-developed approaches improve both efficiency (achieving certain levels of measurement precision with fewer items) and precision (more accurately estimating true theta), with generative AI-based approaches demonstrating the strongest performance. Future work will validate these results using data from an upcoming collection involving human US college students.

Mayank Kumar

Enhancing test creation using Automated Item Generation

Abstract: One of the challenges for organizations that conduct exams on regular basis is availability of items. Generating items regularly is a labor intensive and time-consuming process. A potential solution to this is Automated Item Generation (AIG). With AIG one can generate large number of items in a very short duration with almost zero cost. It also supports creation of item banks.
The model which we have developed leverages pre-defined customizable templates along with Natural Language Processing (NLP) to generate questions based on Quantitative aptitude and logical reasoning. The reason we went for template-based model instead of deep learning algorithms is that deep learning algorithms are fairly complex and require high end machines to operate. Moreover, it is easy to control and minimize errors in template-based models. Some of the examples of items generated through AIG model was computation, arithmetic problems, quadratic equations, inequalities, syllogism, logical reasoning puzzle, coding decoding, etc. The effectiveness of the model was evaluated by giving it to reviewers without telling them whether the items were generated by humans or by machines. The items generated were later on tested on a large population. In order to make the machine generated items more dynamic, for certain type of items we are moving towards Large Language Models (LLMs).
Our finding showed that reviewers were unable to classify whether the items were generated by humans or by machines. The post exam analysis showed that the statistics such as omission rate, mean score of AIG items were similar to that of human written items .Moreover, template-based AIG model improved the quality and diversity of test banks while also cutting down on the time and expense involved in creating test items manually. In future, we intend to explore LLMs for generating arithmetic items.

S1-2: Paper Session - CAT Applications 1

Chair: Yeonwho Kim

Yeonwho Kim & Hyo Jeong Shin

Recommendations for reporting practices: comparisons of proportion correct between linear tests and the adaptive tests using the PISA 2022 data

Abstract: Recently, many large-scale assessments have shifted to adaptive testing, which uses a test-taker's interim performance during the test to improve measurement efficiency and enhance the test-taking experience. More specifically, multistage adaptive testing (MST) designs are widely used, in which subsequent modules (i.e., item groups) are selected according to the test-taker's performance on previous modules. Some notable examples are the MST designs of PISA 2018 Reading and PISA 2022 Mathematics and Reading.

As the large-scale assessment data from the MST designs have begun to be collected, it is appropriate to examine whether the traditional statistical procedures and methods are still applicable. For example, regarding the application of item response theory (IRT)-based measurement models, Mislevy and Chang (2000) pointed out that the essential conditional independence assumption is not robust to the data collected from adaptive tests, and Zhang (2013) improved the dimensionality analysis method to deal with the adaptive test data.

More recently, Ali et al. (2024) conducted simulation studies and showed that basic item analysis statistics based on Classical Test Theory (CTT), such as proportion correct and R-biserials, are biased for the data collected from adaptive testing. That is, when calculating the proportion correct for the given item, a lower-than-expected value may be returned if a group of test takers is low performing. Conversely, a higher-than-expected value may be returned if a group of test takers is high performing.

The problem is that these CTT-based statistics may still be computed by many testing practitioners to check item quality for MST data because these quantities are easy and straightforward to compute and interpret. In this context, it is particularly important to illustrate the misbehavior of CTT-based statistics for the data collected from the MST designs for scientific reporting purposes. To address this challenge, PISA 2018 introduced and operationalized the use of the "equated P+" method, which adjusts the conventional proportion correct by incorporating IRT models.

In this study, we aim to illustrate the misbehavior of CTT-based statistics when data are collected under the MST designs and provide recommendations for reporting practices to deal with such data. Similar to the equated P+ methods proposed in PISA 2018, our recommendation would be based on the use of IRT models but applied to a single group of test-takers. More specifically, we focus on the behavior of the proportion correct, as this is the quantity that is predominantly used in real testing practices. We begin by replicating the simulation studies reported by Ali et al. (2014) to validate our analytical tools, and then illustrate the differences in proportion correct between the linear and adaptive tests using the PISA 2022 Mathematics data collected in Korea. By adjusting the proportion correct using the IRT methods,

we recommend the use of IRT-based adjustments for more scientific and sounder reporting practices.

References

Ali, U., Shin, H., & van Rijn. (2024). Applicability of Traditional Statistical Methods to Multistage Test Data.

Mislevy, R. J., & Chang, H. H. (2000). Does adaptive testing violate local independence? *Psychometrika, 65*(2), 149-156.

Organisation for Economic Co-operation and Development. (2024). P*ISA 2022 technical report*. PISA: OECD Publishing.

Zhang, J. (2013). A procedure for dimensionality analyses of response data from various test designs. *Psychometrika, 78*(1), 37-58.

Dong Gi Seo, Jeongwook Choi, & Dogyeong Kim

The Impact of Differential Item Functioning on Ability Estimation using Korean Medical License Examination within Computerized Adaptive Testing

Abstract: This study explores the impact of Differential Item Functioning (DIF) on ability estimation within the context of computerized adaptive testing (CAT), utilizing the real data from the 2017 Korean Medical License Examination (KMLE). Despite the growing application of CAT, limited methodologies have been specifically tailored for detecting DIF in such assessments. Historical studies have predominantly relied on simulations to analyze DIF effects, leaving a gap in research conducted with actual examination data. This research fills that void by incorporating real exam data to better understand how DIF influences ability scores in CAT settings. The KMLE utilized in this study involved 3,264 examinees and 360 items, spread across eight distinct content areas, each with unique learning objectives. In simulating a realistic CAT environment, the study employed a CAT algorithm that initially randomly selected five items from any content area. Subsequently, to align with the predetermined test plan, the content area least represented relative to its target percentage was chosen. Item selection within the content area was then based on achieving a target response probability of 60%, aiming for desired content coverage specified in the KMLE test plan. Administered in five different testing centers in Korea, the KMLE was designed as a paper-based test, requiring approximately four hours for 2,072 male examinees and 1,192 female examinees. The test comprised multiple-choice questions with one correct answer, with responses recorded as binary outcomes (0 or 1). This study was designed to compare with accuracy and efficiency of ability by Gender in the two item banks in CAT with respect to DIF study. First, item parameters with 96 items of Gender DIF are previously stored in an item bank including all 360 items, facilitating their use in this post-hoc CAT simulation. Second, item parameters without 96 items of Gender DIF are previously updated in the item bank including 264 items. Two post-hoc simulation designs enabled the examination of Gender DIF items' role in estimating examinee abilities, particularly focusing on CAT test conditions.

In order to evaluate the estimation results, three indices are calculated; bias, Root Mean Squared Error(RMSE) and the mean of number of items administered. The bias indicates a difference between the true abilities and the estimated abilities. The bias, RMSE, the mean of administer items between the bank with the DIF items and the bank without the DIF items will be compared in post-hoc CAT simulation under different true ability conditions. The findings are expected to reveal that the existence of DIF items significantly affects the accuracy of ability estimates with more substantial bias. This bias is anticipated to diminish when DIF items are not existed in the assessment. The analysis also predicts a pronounced impact of DIF on examinees at the extremes of the ability spectrum, with the most skilled individuals likely experiencing the greatest underestimation of their abilities. These insights highlight the importance of strategic DIF item placement and detection in enhancing the fairness and precision of CAT-based assessments.

Jungsoo Kim, Jeongwook Choi, & Dong Gi Seo

**Application of Computerized Adaptive Testing (CAT) to vocational fundamental
competency Test**

Abstract:

Purpose: With the government's introduction of blind recruitment, National Competency Standards (NCS) have become a standard assessment tool in South Korea. However, the validity of these paper and pencil test has not been thoroughly researched. The COVID-19 pandemic accelerated the computerization of exams, making it necessary to consider CAT in the vocational fundamental competency. This study examines the feasibility of applying CAT using simulation studies

Method: In this study, in order to investigate the possibility of applying computer adaptive testing (CAT) to vocational fundamental competency tests, we used Monte Carlo simulation to generate response data for s and investigate the accuracy and efficiency of CAT according to differences in test taker levels. For this purpose, the true abilities of 3,000 people were randomly sampled from a normal distribution with means of -1, 0, and 1, respectively, and standard deviation of 1. The research methodology involved several key steps. The bank consisted of 242 s used in actual recruitment. To estimate the difficulty of the response data were collected from job seekers and college students through an online mock test, and parameters were estimated using Winsteps 4.0.1 based on the Rasch model. The starting rule was set to randomly provide the initial five s regardless of difficulty, and the Maximum Fisher Information (MFI) criterion was used as the selection rule. Three methods were used to estimate candidate ability: Maximum Likelihood Estimation (MLE), Expected a Posteriori (EAP), and Maximum a Posteriori (MAP). Standard Error of Estimation (SEE) was used as the termination rule, with thresholds set at 0.3, 0.25, and 0.2. The accuracy of the CAT was assessed using bias, root mean square error (RMSE), and efficiency based on the number of s administered.

Results: In this study, we found that all three estimation methods used (MLE, EAP, MAP) provide satisfactory accuracy. There has been no significant reduction in the number of items offered through CAT compared to the number of items provided in current exams. However, fairness has been improved so that measurements and evaluations can be made with the same accuracy. In particular, it is expected that fairness can be secured without a significant change in the number of items in the current test, thereby reducing confusion among test takers due to changes in testing tools.

Conclusion: This study concluded that although CAT did not significantly reduce the number of items in the current vocational fundamental competency Test, this is because the number of items in the current vocational fundamental competency Test is not large, and that the fairness of the test can be improved through CAT. This study serves as a preliminary simulation study to explore the feasibility of CAT application, indicating that additional post-hoc simulation studies and LIVECAT studies are necessary for

actual implementation. Furthermore, the study emphasizes the need for content balancing in future CAT implementations and suggests further research incorporating content balancing constraints.

S1-3: Paper Session - Advancements in Adaptive Testing

Chair: *Alina Von Davier*

Alina Von Davier

Advancements in Adaptive Testing: Introducing Personalized Ensemble Tests (PETs)

Abstract: Achieving a robust match between test items and test takers' abilities is crucial for effective testing instruments. The need for adaptation arises when addressing the heterogeneity within the test-taking population. Our prior work has led to the development of hybrid testing approaches, such as Computerized Adaptive Tests (CATs), Multistage Tests (MSTs), and random item assignment, aimed at optimizing this match. However, we now aim to refine this match further by personalizing not only the difficulty level of the content but also the content itself to align with test takers' individual characteristics and testing rationale. Leveraging advancements in technology, particularly AI and ML techniques, we introduce a novel test design termed Personalized Ensemble Test (PET). Under this framework, a CAT is a PET, (and so is an MST) offering a more tailored and effective testing experience.

Furthermore, recent strides in AI for item generation, coupled with new adaptive algorithms borrowed from machine learning, have significantly enhanced the viability of such a complex test design. The availability of vast computational power also supports the implementation of a Bayesian approach to pool calibration, further refining the accuracy and reliability of PETs.

These technological advancements open new avenues for personalized testing, ensuring a more comprehensive evaluation of test takers' abilities while maintaining efficiency and accuracy.

Angela J. Verschoor

Accuracy of short variable-length CATs

Abstract: Main advantage of a computerized adaptive test (CAT) is its improved efficiency, giving us the possibility to reduce test length while maintaining a desired level of accuracy. A logical strategy would be to use the standard error of measurement (SEM) as a stopping criterion in a variable-length CAT. Yet, it is known that short variable-length CATs are problematic: Babcock and Weiss (2012) investigated the matter, coming to the conclusion that only at lengths of at least 10-15 items, the SEM can be evaluated accurate enough to serve as a stopping criterion.
This study concentrates on the factors that cause the relative inaccurate measures of SEM in short-length CATs. Obviously, each ability estimator provides not only an estimation, but also an estimated SEM. This leads us to our research question: which estimator provides the most accurate SEM, especially in the case of short tests? The SEM will generally be large in short tests, but question here is which estimator gives the most accurate estimation of SEM. A second question, related to the first, is if we can identify factors that cause the estimated SEMs to be inaccurate.
Three different estimators were considered in this study: MLE, WLE and EAP. While MLE and WLE share a certain similarity maximizing (a function of) the likelihood, EAP takes a Bayesian approach whereby the estimated SEM is formed by the standard deviation of the posterior distribution. A simulation study using an operational CAT item pool was performed to make a mapping of the estimated SEM of the three estimators. In the first phase, a set of simulated students with various, fixed thetas took CATs of fixed length. A sample size of 10,000 per theta point was chosen, each run replicated 100 times. The simulations showed that the most likely cause of the inaccuracy of the estimations of SEM is the occurrence of zero and perfect scores, and how the three estimators react on those scores. The second phase of the study was a simulation of short variable-length CATs. As zero and perfect scores tend to be associated with large SEMs, these scores tend to occur less frequently than in fixed-length CATs or linear tests of on average the same test length. As a result of this, the inaccuracy of the estimations of SEM tends to be significantly reduced.
The WLE produced the most accurate estimations of SEM in all conditions, while the performance of MLE in this respect depended entirely on the somewhat ad-hoc methods of truncation or fencing in the case of zero and perfect scores. EAP tended to overestimate its SEM, showing a significantly lower error variance than the variance of the posterior distribution. Using the WLE will alleviate the problem of inaccurate estimations of SEM to manageable proportions, thus giving a justification for variable CATs with lengths of approximately 3 items or more.

Reference

Babcock, B. & Weiss, D. (2012), Termination criteria in computerized adaptive tests: Do variable-length CATs provide efficient and effective measurement?, Journal of Computerized Adaptive Testing, 1 (1), 1-18

Vipin K Chilana, Mayank Kumar & Jyoti Pandey

Successfully Implementing CAT without sharing the right answer key in the Delivery System

An immediate reaction to such a proposal may be of utter disbelief. One reaction may be it is impossible? Another would be "Why, is there a need for it?"

Background: Generally, Test Content along with Right Answer Key (Key) together are considered a part of content security. This usually is the case. Compromise with content itself without the associated key is a major breach. But when breached with the associated key it becomes an easy source of manipulation. Some cases of manipulation only on the basis of key (without accessing content) have also come to the light. However, if content and key can be segregated it adds an added layer of security. In this model, key and content are not only segregated and unbundled, key is taken out of the equation and is not at all required in delivery of the examination content. It has been experimented and proved that in some cases where Content-key compromise was suspected it could be foiled because the actual key was not provided in the exam delivery system. IBPS has been into Computer Based Tests (Classical Test Theory based) since 2006 and in 2006 it was devised to deliver exams at the centers without the key. For Classical Test Theory based examinations, where the result/score was not to be given immediately to the candidates, it, in retrospect now, looks a very simple solution. But till 2006, till we implemented it, nobody else had gone in that direction. Later some other testing organizations/Test Delivery partners in India have started emulating it.

India and some of the Asians markets present unique and varied set of complexities in which content security and attempts of malpractices and unfair means related to content assume an important role. Most of the exams are delivered in Local Area Network where content is downloaded in the Local Server Machines. However, conceptualizing without key for adaptive models, where next question depends on whether the answer to the previous question is right or not, posed problem even at deep thinking stage. Thought Experiment is this case proved to be a tremendous problem solving tool.

In India, High Stake Exams are delivered in the following two major modes

IBA – Internet Based Assessment – Exams are delivered from a central server and test takers are connected to it through internet at proctored Test centers. Security of content and key is not so high as it resides only at the Central Server. Linear, Adaptive and MST mode can be implemented. However, scaling of numbers in a session poses problem. Over-dependence on internet is an impediment resulting in disruptions and delay. Scaling up beyond about 40,000 a session is a challenge.

LBA – Lan Based Assessment – Each Centre of about 250 candidates/examinees has its own server for exam delivery and test takers are connected to it through Local Area Network (LAN). The Content is

downloaded on each server in a secure manner and delivered to examinee's machine also in a secure manner. Scaling up numbers per session in LBA is not limited by internet connection and server concurrency and is restricted only to the extent of available Computer infrastructure. The model has least dependence of internet connectivity during the examination. So far about 200,000 test takers per session can be accommodated. This number can be further scaled up.

Though the "without Key" model was reimagined in LBA mode it can also be applied in IBA mode. Though the solutions slightly differ but can be applied in both Item level adaptive as well as Testlet level (Adapting within a testlet) with the similar kind of solution. For Multistage Adaptive Tests (MST) it requires a slightly different approach.

Kristof Kovacs, Hanif Akhtar, & Balazs Klein

Computerized adaptive testing made enjoyable: finding the optimal trade/off between measurement efficiency and user experience

Abstract: Since in computerized adaptive testing (CAT) the difficulty of the selected items approximates the test-takers' estimated ability, it has often been claimed that CAT results in a better test-taking experience than traditional tests. Our recent meta-analysis, however, did not provide universal support for this claim: motivation was not univocally higher and/or anxiety lower when people were administered a CAT instead of a traditional fixed-item test (FIT). The only exception was when an easier version of CAT (ECAT) was administered, i.e., a CAT algorithm was modified to select items with difficulty lower than the current ability estimate so that participants had a mean success rate higher than 50%. According to a limited number of studies ECAT indeed seems to have positively affected examinees' testtaking experience. At the same time, from a psychometric perspective it can be argued that ECAT is inferior to CAT as it results in decreased measurement efficiency: under ECAT either the error margin of the ability estimate is higher than under CAT or more items have to be administered in order to arrive at the same accuracy.

The purpose of the present study was to comparatively investigate both the psychometric and psychological effects of traditional CAT, ECAT, and fixed-item testing (FIT). We tested a sample of junior high school students (N = 428) under a randomized experimental design: participants were randomly and blindly assigned to one of the three groups (CAT, FIT, ECAT).

ECAT was superior to CAT and FIT in terms of its psychological effects: it resulted in lower anxiety, higher effort, and greater perceived performance. As expected, the same number of items administered in all three groups resulted in a lower measurement precision for ECAT than for CAT, even though ECAT was still superior to FIT in this respect. However, we found that participants in the ECAT group solved items much faster: on average, completing the ECAT required 35% less time than regular CAT. That is, since test-takers are much faster when completing the easier items administered under ECAT than participants in the CAT group who are presented with regularly selected items, the additional items that need to be administered under ECAT to achieve the same accuracy as under CAT do not result in increased testing time and in turn in increased fatigue.

Overall, this study demonstrates that modifying the CAT item selection algorithm to select easier items results in a better user experience without sacrificing measurement efficiency. Under ECAT, in order the achieve the same precision a larger number of items need to be administered, but the length of the ECAT session still does not exceed the length of the CAT session. Therefore, ECAT can be a favorable method of computerized adaptive testing, especially when assessing children.

S1-4: Symposium - Advancements in Bayesian Methods for CAT

Chair: *Jonathan Templin & Andreas Frey*

Abstract: This symposium brings together a group of researchers who work independently but on a similar topic: the use of Bayesian methods in computerized adaptive testing (CAT). The symposium provides a series of studies that may extend the capabilities of CAT through the integration of Bayesian methods. These contributions collectively address several limitations in traditional CAT by enhancing the precision and reliability of the testing process and ensuring that bias due to estimation or sampling error is minimized.

The focus of this symposium is on exploring novel Bayesian techniques that mitigate issues such as the underestimation of item parameter uncertainty, problems due to calibration error, the inefficiencies of fixed test lengths, and the challenges posed by small sample sizes in item calibration. The four presentations cover a range of approaches, from fully Bayesian adaptive testing with variable-length stopping criteria to the integration of Bayesian explanatory models in diagnostic classification and item response theory. Each study provides a glimpse into how Bayesian methods were systematically applied in differing testing situations, each of which provides some insights into the impact Bayesian methods could have on both the theoretical underpinnings and practical implementations of CAT.

The results of each of these studies show how differing methods can improve not only the accuracy of ability estimation and the effectiveness of item selection but also aim to set new standards for conducting adaptive testing in educational, psychological, and other formative assessment environments. By embracing Bayesian methods, the symposium highlights innovative solutions to enhance the fairness and efficacy of adaptive testing practices, ensuring that test outcomes are both scientifically robust and practically relevant. Moreover, the methods presented in this symposium could be useful for future advances in automatic item generation, enabling CAT with on-the-fly item generation.

The symposium is designed for academics, practitioners, and policymakers interested in the latest developments in adaptive testing technology. Attendees will gain insights into the potential of Bayesian methods to empower CAT applications, driving the evolution of adaptive assessments to meet the diverse needs of modern learners and educational systems.

Presenters:

Andreas Frey & Aron Fink

Accounting for Item Calibration Error with Multidimensional Bayesian Adaptive Testing

Abstract: conceptual shortcut that makes CAT easier to handle but can lead to substantial problems such as underestimated standard errors and bias in ability estimates. This simplifying shortcut is that, in item selection and ability estimation in CAT, the point estimates of item parameters that were estimated in the calibration study (and thus contain measurement error to a certain degree) are used as fixed values that do not contain any measurement error. This problem is especially pronounced in multidimensional CAT (MCAT), where multiple abilities are assessed simultaneously and, therefore, the accuracies of the resulting ability estimates for multiple dimensions are affected concurrently. In order to address this fundamental problem, we propose a novel Bayesian algorithm for MCAT that explicitly models item parameter uncertainty during item selection and ability estimation. This method uses Bayesian updating for item and ability distributions. The item parameters are not introduced as fixed values when the multivariate posterior ability distribution $\boldsymbol{\theta}_{j}^{(t)}$ of person j is updated but, instead, are sampled from stored post burn-in draws from an MCMC-based calibration. Similarly, the post burn-in draws for the previous multivariate ability estimate $\boldsymbol{\theta}_{j}^{(t-1)}$ serve as the prior distribution of the next ability estimate after a new response is registered. This method avoids the abovementioned shortcut. In a comprehensive simulation study, the performance of the novel method is compared with conventional MCAT approaches. A multidimensional adaptive test with three dimensions mutually correlated with .50 is used in all conditions. The multidimensional 2PL model is used as psychometric model with a mixture of between- and within-item multidimensionality. The simulated item pool comprises 300 items, and the test length is 60 items. The simulation study is based on a factorial design with the factors calibration sample size (250; 500; 1,000), method for accounting for uncertainty in item parameter estimates (none, Bayesian), and true ability per dimension $\theta = (-2.0, -1.5, \cdots, 1.5, 2.0)$. The evaluation criteria are bias and MSE of the three-dimensional ability estimates, each conditional on the true ability levels of the test takers. The simulation is carried out with R and Stan is used for the MCMC estimation. Although the simulation is ongoing, preliminary results suggest that the fully Bayesian approach leads to less bias as well as to lower MSE values compared to conventional MCAT methods, especially for high- and low-ability levels. The final results will be presented at the conference. To conclude, the suggested Bayesian method seems to be a promising method for making MCAT more accurate, which is a prerequisite of unbiased ability estimates with correct standard errors.

Luping Niu & Seung W. Choi

Variable-Length Fully Bayesian Adaptive Testing and its Stopping Criteria

Abstract: Many computerized adaptive testing (CAT) programs currently use fixed item parameters based on point estimates obtained during item pool calibration. However, ignoring the uncertainty in these estimates can lead to underestimating the standard error of trait levels. To address this issue, van der Linden and Ren (2020) introduced a novel approach to a Fully Bayesian (FB) CAT algorithm. This algorithm integrates the uncertainty of item parameter estimates into trait estimation, a feature that sets it apart from the conventional CAT under fixed-length conditions. However, a closer examination under variable-length conditions is crucial to gain deeper insights into the differences between the FB and conventional CAT approaches. Specifically, underestimating the standard error in variable-length conditions can cause premature terminations and directly impact the fidelity of termination decisions. The current study investigated the FB CAT algorithm in variable-length CAT, where examinees receive differing test lengths based on preset termination criteria. Additionally, the study proposed various stopping criteria suited to the FB algorithm, including the posterior standard error (SE), change-in-theta estimates (CIT), predicted standard error reduction (PSER), and the combination of CIT and SE. For the SE-based stopping rules (i.e., SE and PSER), the FB version of SE of theta used in the variable-length CAT can be obtained by directly calculating the standard deviation of the posterior sample theta. Likewise, the FB adaptation of the CIT criterion can be straightforward. However, the FB adaptation of the PSER stopping rule may require a separate MCMC sample for each predicted response (e.g., 0 vs. 1) for all available items, which can be computationally expensive for real-time selection of items. The study delineates the necessary adaptations to FB CAT and potential simplifications of the stopping criteria. The study then compared the effectiveness of these adapted FB stopping criteria regarding test length and estimation accuracy. Finally, the study identified the most effective stopping criterion for the FB CAT algorithm based on extensive simulation results.
Using the SE stopping rule, the FB algorithm demonstrated superior performance to the conventional counterpart in estimation accuracy in variable-length CAT, mainly when the calibration sample size was small. The CIT rule produced a more consistent test length across different theta levels than the SE stopping rule but required considerably more items than the PSER rule. The PSER and a combination of CIT and SE rules were superior to others in balancing estimation accuracy and test efficiency, especially with larger calibration samples. These rules consider the posterior SE and, at the same time, incorporate predicted SE reduction or theta change, allowing the exam to be terminated if the remaining items are unlikely to contribute further to SE reduction or theta change.

Concerning computational speed, the FB CAT algorithm with SE, CIT, or combined CIT-SE stopping rules was extremely fast, averaging 0.03 seconds per item, making it ideal for high-volume CAT applications. The PSER stopping rule, while slower than other criteria, remains competitive and close to real-world CAT application standards, given its top performance in the FB variable-length CAT framework. It is important to note that selecting the optimal stopping rule ultimately depends on the specific needs of the testing program and the desired trade-off between accuracy and efficiency. These findings have significant implications for the design and implementation of CAT systems, providing a clear roadmap for researchers and professionals in the field.

Ae Kyong Jung, Jonathan Templin, & Nathan DePuy

Integrating Bayesian Explanatory DCMs and Item Selection Algorithms for Small Sample CAT

Abstract: Our study investigates the combination of Bayesian explanatory diagnostic classification models (DCM) and novel item selection algorithms to address the challenges arising from limited sample sizes in computerized adaptive testing (CAT). Typical CAT algorithms, which rely on point estimates of item parameters, have been shown to have a decrease in classification accuracy, which effects the accuracy of measurement in DCM due to calibration errors. Proposed methods to mitigate these effects include improving item quality, using large calibration samples, and extending the test length (e.g., Huang, 2018; Sun et al., 2020). However, Bayesian estimation for small sample settings has not been thoroughly studied. Therefore, our study aims to study the impact of Bayesian Explanatory DCMs on the precision of item parameter and examinee profiles.

Bayesian estimation enables the mitigation of uncertainty by sampling from the posterior distribution of all model parameters for item selection and ability estimation (e.g., Ren et al., 2020). This study introduces Bayesian item selection algorithms to effectively employ these posterior distributions of parameters in CAT. Our item selection methodology, which is based on Shannon entropy (Shannon, 1948), the uncertainty criteria for discrete latent variable, involves drawing samples from the posterior distribution of the item and examinee profiles. Items are administered by selecting the next item with the lowest value average Shannon Entropy. This approach enables carryover of uncertainty associated with all model parameters, which leads to better item exposure rates, particularly in small samples. In this study, we compare a new algorithm with conditions that employ point estimates such as expected a posteriori (EAP) or maximum a posteriori (MAP) of posterior distribution to compare with traditional CAT based on maximum likelihood estimation. The same sampling process is applied when estimating the examinee profiles.

To increase the precision of estimates calibrated with small samples, our approach incorporates an explanatory DCM into the algorithm. In the explanatory DCM, item parameters are regressed onto a set of item properties rather than estimated separately. As item calibration for large item pools necessitates a large sample size per item to derive item parameters (e.g., Segall, 2005), the explanatory DCM efficiently estimates item properties instead of individual item parameters, owing to the fewer number of item properties compared to items. This reduces estimation complexity compared to traditional models, which is beneficial to small sample calibration (e.g., Wilson et al., 2008). To evaluate the effectiveness of these algorithms, our study encompasses both a simulation study and an empirical data application.

Simulation Studies:

The simulation studies start by generating an initial pilot sample (N=30, 300, or 2000). The true examinee profiles are randomly assigned so that each examinee has a 50% chance of mastering each attribute. The responses are generated using item parameter values from a pilot study on a new assessment of English morphology in elementary grades students. Item calibration for the item pools uses Gibbs sampling algorithms. Item selection criteria based on Shannon entropy are calculated by four conditions (EAP, MAP, sampling 1 draw, and sampling multiple draws from the posterior distribution). The posterior distribution of examinee profiles is updated also using the same four conditions with item selection criteria. The adaptive algorithms stop when the mean of posterior probability of examinee's profile exceed .7 or .8. After obtaining responses from a new set of students, the item pool is then recalibrated, and updated parameters are used for the next students.

Jonathan Templin, Ae Kyong Jung, & Nathan DePuy

Computerized Adaptive Testing Using Bayesian Explanatory Item Response Theory

Abstract: This study expands upon the integration of Bayesian explanatory item response models (EIRT) with computerized adaptive testing (CAT) algorithms, focusing on enhancing precision in assessments with small sample sizes with ramifications for evaluating the use of EIRT for automatically generated items. This research was initiated due to challenges faced during a 2011 pilot study funded by an NIH-supported grant, which aimed to develop a CAT for a visual attention task involving on-the-fly item generation and the use of EIRT models. Drawing from experiences where traditional CAT methodologies fell short, particularly in handling small datasets, this study seeks to address these gaps by introducing innovative Bayesian algorithms for item selection and ability estimation.

The pilot study data, which informs this research, originated from a project on how students in grades 3-5 learn English morphology rules. The grant, received just before the onset of the COVID-19 pandemic, was constrained by the limited availability of large student samples. This challenge necessitated the development of creative methods to integrate CAT into the assessment, given the restricted data collection opportunities. The methods described herein are derivatives of the lead author's initial consultancy on the NIH grant, where the necessity for adaptive assessment tools capable of handling sparse data was first recognized.

Our approach employs Bayesian algorithms that adjust for the common underestimation of standard errors in ability estimates traditionally seen with the reliance on point estimates of item parameters. We utilize the posterior distribution of model parameters for both item selection and ability estimation, enhancing precision by robustly accounting for parameter uncertainty. Our item selection methodology, in particular, involves calculating item information by drawing multiple samples from the posterior distribution of item and ability parameters. Items are then selected based on the highest mean information value, thereby adjusting for uncertainty across all model parameters.

Moreover, we integrate an explanatory IRT model that models student responses based on item properties rather than unique parameters per item. This adaptation is crucial for efficiently managing the complexity of assessments and allows for an in-depth examination of how specific item properties influence item difficulty. Such a model is especially beneficial in scenarios like ours where item calibration for large item pools would typically require a large sample size per item to derive accurate parameters. Additional potential benefits include the ability to incorporate on-the-fly items from automatic item generation methods directly into operational computerized adaptive assessments.

To further validate our methods, simulation studies were conducted with varying item pool sizes, sample sizes, and selection functions.

These studies utilized the pilot data from the English morphology assessment, followed by recalibration after collecting responses from new student groups. By manipulating factors such as item pool size (30, 99), initial pilot sample size (30, 300, 1000), incremental sample size (30, 300), item selection function (sample 1, sample 10, sample 100, EAP, median, mode), item summary function for ability estimation (sample 1, sample 100, EAP), and item inclusion after sampling (30, 99), we sought to thoroughly evaluate the effectiveness of our proposed Bayesian CAT methodologies.

This investigation provides a robust framework for enhancing the reliability and validity of assessments, particularly in educational settings challenged by limited data availability. Through the use of Bayesian methods and explanatory IRT models, this study not only addresses the immediate needs of the pilot project but also sets a precedent for future assessments dealing with similar constraints.

S2-1: Symposium - Research for Practical Issues and Solutions in Computerized MST

Chair: *Duanli Yan*

Abstract: Artificial Intelligence (AI) has become a buzzword and receives a lot of attention, but much of this focus is on large language models. Often overlooked is that adaptive/multistage testing (MST) is a form of AI, and our field has been a leader in AI research since the 1960s, such as the algorithms for merging adaptive learning and adaptive testing (Ferguson, 1969).

Multistage testing provides valuable information about individuals, enabling decisions such as hiring, professional skills certification, university admissions, medical diagnosis, or determining the next teaching step for a 4th-grade math student. We aim for these tests to provide accurate information as efficiently as possible. MST maximizes the information gained per minute of testing. By administering items that match the K-12 student's level, we can more effectively determine their mastery level, reducing time away from actual learning, increasing test security, and enhancing engagement by tailoring difficulty to the examinee's ability.

With the rapid increase in large-scale MST applications, many testing institutions and researchers have gained substantial experience in addressing operational challenges and practical problems. This symposium showcases recent research and applications in MST, marking the debut of the new book Research on Practical Issues and Solutions in Computerized Multistage Testing (2024). This book provides the latest research on various practical considerations, methodological approaches, and solutions for implementing MST in operational applications, such as the 2024 Scholastic Aptitude Test (SAT). It builds on the legacy of earlier works, such as *Computerized Multistage Testing: Theory and Applications* by Yan, von Davier, and Lewis (2014). The presentations are authored by renowned psychometricians recognized globally for their expertise in adaptive and multistage testing.

This symposium features six presentations:

1. **Background and Adaptive Testing in the Current AI Age** - an overview of adaptive testing's evolution.
2. **History and Introduction of Adaptive and Multistage Testing** - a look at the historical development and current practices in adaptive and MST.
3. **Designing MST to Meet Accuracy and Efficiency Goals** - considerations for designing MST to optimize accuracy and efficiency.
4. **Innovative Approaches for MST Routing** - new methods for routing in MST.
5. **Development and Applications of Probability** - Based Classification for MST- exploring probability-weighted classification methods.
6. **AI's Impact on Adaptive Testing** - discussing how AI advancements will influence adaptive testing.

These presentations address the latest research in adaptive and MST, tackling challenges and critical issues from theory to practice and showcasing innovative approaches and solutions developed by practitioners.

Presenters:

Nathan Thompson

Foreword

Abstract: The concept of adaptive testing has long been integral to education, as evidenced by the teacher who quizzes a student and assigns subsequent readings based on their performance. Adaptivity is a natural approach to assessment, learning, and many other aspects of life, from athletics to movie recommendations on Netflix. his is an exciting time for research in MST and other AI applications in psychometrics. The pace of research has accelerated with greater access to journals, books, and open-source software like R. The market's expansion, with more commercial solutions available at lower costs, makes MST feasible even for small organizations. Technological advancements continue to enhance item quality and computational capabilities, enabling tasks like automated essay scoring, adaptive/multistage simulations, and process data analysis.

Dave Weiss & Duanli Yan

A Brief History of Computerized Adaptive and Multistage Testing

Abstract: Adaptive testing tailors test item presentation to the individual being tested, aiming to measure each person as precisely as possible while minimizing test length. Adaptive tests administer items, score responses, and select new items based on those scores, continuing until the test meets its objectives. Though often applied in educational contexts, adaptive testing has broader applications that have shaped contemporary computerized adaptive testing (CAT) and MST. This presentation reviews the history of adaptive testing and introduces current practical issues and solutions focusing on stage-based adaptive testing.

Mark Reckase, Unhee Ju, & Sewon Kim

Designing Multistage Tests to Meet Accuracy and Efficiency Goals

Abstract: MSTs offer advantages over item-level CAT by allowing tighter control of test content and item review within modules. However, there is limited research linking MST configurations to specific measurement goals. This presentation discusses designing MSTs to achieve equal module usage and uniform test information over a proficiency range, illustrating well-designed item pools and their effectiveness.

Kyung T. Han

Multistage Testing with Inter-Sectional Routing for Short-Length Tests

Abstract: MST has advantages over item-level CAT but offers reduced adaptability. Typically, the first stage is a routing stage where all examinees see a linear test form. This presentation

proposes a new framework for MST with Inter-Sectional Routing (ISR), evaluated under various conditions. The findings suggest that MST with ISR improves measurement efficiency and test optimality, especially for short-length tests.

Victoria Song, Duanli Yan, & Charles Lewis

Development and Application of Probability-Weighted Classification for Multistage Testing

Abstract: Classification testing sorts test-takers into groups based on scores. This process traces back to Wald's Sequential Probability Ratio Test (SPRT) in 1947. Modern applications include adaptive mastery testing and MST. This presentation introduces mathematical models using probability weights for binary classification into accepted or rejected categories.

Alina A. von Davier

How Will AI Change Adaptive Testing?

Abstract: Recent advances in AI, ML, big data management, and computational power have transformed various industries, including education. This presentation explores the implications of generative AI for adaptive testing, discussing what will change and what will remain the same, with examples of current applications.

S2-2: Paper Session - Item selection methods 1

Chair: *James Sharpnack*

James Sharpnack, Kevin Hao, J.R. Lockwood, Steven Nydick, Alina A. von Davier

A Thompson Sampling Approach to IRT-based Computerized Adaptive Tests

Abstract: We present two new methods for item selection in computerized adaptive testing (CAT), an adaptive selection algorithm called BanditCAT, and an extension called S2A3. We motivate our approach by casting the problem in the contextual bandit framework and Bayesian Item Response Theory (IRT). Contextual bandits are machine learning (ML) methods that use a reinforcement learning approach for sequential decision making, and have been used in recommender systems across industries. Here the key insight lies in defining the bandit reward as the Fisher information for the selected item, given the latent test taker ability (θ) from IRT assumptions. We use Thompson sampling—an approach commonly used in contextual bandit literature—to balance between exploring items with different psychometric characteristics and selecting highly discriminative items that give more precise information about θ.

As an extension of this, we introduce Soft-Scoring and Adaptive Adaptive Administration (S2A3), which accounts for uncertainty in item parameter estimates and enables periodic item re-calibration. Soft-scoring (S2) refers to using the Bayesian marginal likelihood when estimating the latent θ, which makes scoring more "fair" by implicitly down-weighting an item's contribution when it has greater IRT parameter posterior variability. Adaptive adaptive administration (A3) utilizes the same Thompson sampling approach in BanditCAT, but adds an additional draw from the item parameter posterior distribution. This methodology enables the scoring and administration of items for which we have little to no response data.

This work bridges the gap between ML approaches to sequential decision making and adaptive testing. We compare and contrast this method to existing approaches including BOBCAT (Gosh et al., 2024) and α-stratification (Chang & Ying, 1999). Through realistic simulations based on data from the Duolingo English Test (DET) and randomized control trials on the DET's practice test, we demonstrate the method's effectiveness, highlighting its potential to enhance test adaptability and precision.

Haejin Kim

Enhancing Adaptive Testing with Golden Ratio Search Method for Personalized Education

Abstract: In the realm of educational measurement, adaptive testing is revolutionizing personalized learning by dynamically adjusting to individual student abilities. This study introduces the application of the Golden Ratio Search (GRS) method for item selection in adaptive tests.
The GRS method optimizes the item selection process by leveraging the mathematical principles of the golden ratio, aiming to enhance both the accuracy and efficiency of student assessments. Traditional adaptive testing methods, like the Maximum Information Criterion (MIC), often require numerous items to precisely determine a student's ability level. In contrast, the GRS method significantly reduces the number of items needed, thereby decreasing test length and minimizing student fatigue.
This research involved a series of simulations comparing the GRS method with conventional item selection techniques. The results demonstrated that the GRS method maintains high precision in ability estimation while enhancing overall test efficiency. Specifically, the GRS method proved effective in accurately assessing student capabilities with fewer test items, making it an ideal tool for educational environments that aim to reduce test-related stress and improve student engagement.
Furthermore, the GRS method's integration into a proprietary educational diagnostic testing service showcases its practical application. The algorithm adjusts in real-time to student responses, providing immediate and tailored feedback that supports personalized learning pathways. This method not only facilitates accurate diagnostics but also aligns educational content to the individual needs of students, promoting a more effective learning experience.
In conclusion, the GRS method represents a significant advancement in adaptive testing, offering a robust, efficient, and student-friendly approach to educational assessments. As adaptive testing continues to evolve, the incorporation of innovative item selection methods like the GRS will be crucial in enhancing the precision and efficiency of personalized education.

Cesar Antonio Chavez Alvarez, Laura Ortega Torres, & Noelia Ramos Casarrubias

An example of applying the Decision Theory to develop Computerized Adaptive Tests

Abstract: In recent years, the benefits of technology have been exploited to create individualized tests delivered by a computer known as Computerized Adaptive Tests. These tests are usually based on models such as Item Response Theory (IRT) or Bayesian, which involve strong theoretical assumptions that require that a large number of people respond to an exam, a requirement that not all tests can fulfill. We evaluate a proposal based on an alternative model that does not require such strong assumptions, known as Decision Theory. This theory makes only one assumption about local item independence. In the study, we took the items and responses of the students who took the Examen General para el Egreso de la Licenciatura (EGEL Plus), a set of university-level tests used as a graduation requisite in Mexico for some courses, developed by the Centro Nacional de Evaluación para la Educación Superior (Ceneval). We used one common area of these tests known as Indirect Writing, a criterion-referenced area designed to assess writing skills with multiple-choice items used to classify students according to their achievement level in these skills, independently of the university course from which they are graduating. We performed a simulation to analyze the hypothetical scenario in which the test was presented to the same students who responded to the real test in its linear form but as an adaptive one based on the decision theory. In the simulated adaptive test we found a high proportion of students classified in the same achievement level as in the real application with the linear form. Although one-third of the simulated population would have responded to the full test, that is 30 items, half of the students would have responded between seven and 16 items. Therefore, Computerized Adaptive Tests based on the Decision Theory may be an alternative to implement an adaptive methodology for criterion-referenced tests. Finally, a comparison of the results of the previous simulation was performed with another simulated more traditional adaptive test based on IRT algorithms.

S2-2: Paper Session – Cognitive Diagnosis CAT 1

Chair: *Junsik Sim*

Junsik Sim

A Comparison of Item Selection Methods in Classification Accuracy and Item Exposure Rate of Cognitive Diagnostic Computerized Adaptive Testing

Abstract: The information provided by testing could not only classify students but also assist learners by providing customized feedback. The Cognitive Diagnosis Model (CDM) offers a multidimensional approach to evaluating examinee ability by identifying whether the examinee has mastered a set of skills. Cognitive Diagnostic Computerized Adaptive Testing (CD-CAT) is a special form of CAT that integrates the cognitive diagnostic model with computerized adaptive testing. However, while CAT can measure an examinee's ability accurately and efficiently, it is essential to ensure that items from the item bank are not overexposed to maintain test security.
The purposes of this study were to identify the differences in classification accuracy and item exposure rates between item selection methods under CD-CAT conditions and to determine the effects of test conditions such as "the number of attributes", "item diagnosticity", and "the number of examinees". To achieve the study's objectives, GDI, PWKL, NPS, and random methods were used as item selection methods in CD-CAT based on the DINA model. To determine classification accuracy, "Pattern-wise Agreement Rate (PAR)" and "Attribute-wise Agreement Rate (AAR)" were calculated. Item exposure rates were presented in frequency tables, and the skewness of item exposure rates was shown as the chi-square (χ^2) of item exposure rates.
The results of this study were as follows: First, classification accuracy was low across the four item selection methods when the number of attributes was large. However, the difference in classification accuracy between the number of attributes in the GDI and NPS methods was not significant when item diagnosticity was high. In terms of item exposure rate, as the number of attributes increased, the χ^2 for the GDI and PWKL methods decreased, while it increased for the NPS methods. When item diagnosticity was categorized into three levels, classification accuracy across the four item selection methods was high when item diagnosticity was high. Under the PWKL method condition, the difference in classification accuracy between high and middle diagnosticity levels was larger than the difference between middle and low levels. The χ^2 of item exposure rates was higher when item diagnosticity was low for the GDI and PWKL methods, but the NPS method showed little difference. When the number of examinees used for estimating item parameters was insufficient, the GDI and PWKL methods imprecisely estimated the attribute patterns, while the NPS methods estimated them regardless of the number of examinees. There was little difference in item exposure rates based on the number of examinees.
In conclusion, the results imply that the GDI method is most appropriate when the number of examinees is large enough. Considering item exposure rates, the NPS method showed the most ideal distribution of item exposure rates. Therefore, the NPS method is most suitable if the number of examinees is insufficient or if the risk of test security is high.

Peng He & Chanho Park

CD-CAT Item Selection Methods Based on Recurrent Neural Network

Abstract: Cognitive diagnostic assessment is attracting attention of researchers in educational assessment since it enables educational treatment for learners by diagnosing their strengths and weaknesses in detail. Cognitive diagnosis with computerized adaptive testing (CD-CAT) is considered a useful and efficient method of implementing cognitive diagnostic assessment in that it allows learners to be diagnosed in real time. In the core of CD-CAT lies the item selection method, which determines test forms and directly affects accuracy and efficiency of measurement. Existing item selection methods, generally based on information theory, use Shannon entropy, mutual information, Kullback-Leibler distance, etc. They require high-quality item banks and accurately calibrated item parameters. They are also dependent on specific cognitive diagnosis models. To overcome this limitation, this study suggests new item selection algorithms based on recurrent neural network (RNN) and its variants, long short-term memory (LSTM) and gated recurrent unit (GRU). These three item selection methods do not require calibrated item parameters and can be applied to various cognitive diagnosis models. The input to the RNN and its variants is the Q-vector of the item and the student's response, and the output is the Qvector of the next item. In this study, the three algorithms—RNN, LSTM, and GRU—are trained using the data generated by the posterior weighted Kullback-Leibler (PWKL) and Shannon entropy (SHE) item selection methods. Monte Carlo simulation analyses are conducted, and the suggested algorithms are compared with the traditional item selection methods such as PWKL and SHE under various condition. The results are evaluated using such statistics as the attribute agreement rate and the pattern-wise agreement rate. Based on the simulation results, strengths, weaknesses, and characteristics of each item selection method are discussed.

S3-1: Paper Session - AIG, Automative Scoring 2

Chair: *Bartosz Kondratek*

Bartosz Kondratek, Margaret Bryndal, & Dan Menzies

Confidence Around the Score in Automated Essay Scoring with Gen-AI

Abstract: Automated Essay Scoring (AES) with generative Large Language Models (LLMs) is one of the many downstream gen-AI tasks actively explored by researchers (e.g., Yancey et al., 2023; Lee et al., 2024). Current research primarily focuses on maximizing the accuracy of these models. To meet the highest standards and regulatory requirements, practical applications of AES adopt the human-in-the-loop (HITL) approach, where human experts monitor the quality of the model-based scores. However, to utilize human expertise efficiently, we need an AES generated score that is not only maximized for accuracy, but also accompanied by a reliable measure of confidence (Funayama et al., 2022).

In contrast to the 'traditional' AES approaches, which all fall under the umbrella of supervised learning, AES with generative LLMs do not come with out-of-the-box means of numerical evaluation of their own output certainty. The present research aims to address this gap by building on two methods suggested in the field of medical diagnosis (Kotelanski et al., 2023):

- Intrinsic Confidence Assessment (IC)
- Self-Consistency Agreement Frequency (SC)

These methods were expanded in two directions:

- IC was 'merged' with SC by directly prompting the generative model to return multiple scores,
- SC was computed over random samples of examples used as the gen-AI's knowledge base.

The quality of confidence measures obtained through the above approaches was evaluated using GPT-4 by comparing the score and confidence pairs to human scores awarded to authentic written responses to examination questions given by English as a Second/Foreign Language learners taking the online, adaptive Kaplan Test of English.

References

Funayama, M., Sato, T., & Takeda, H. (2022). Balancing cost and quality: An exploration of human-in-the-loop frameworks for automated short answer scoring. *arXiv*. https://doi.org/10.48550/arXiv.2206.08288

Kotelanski, M., Gallo, R., Nayak, A., & Savage, T. (2023). Methods to estimate large language model confidence. *arXiv*. https://doi.org/10.48550/arXiv.2312.03733

Lee, S., Cai, Y., Meng, D., Wang, Z., & Wu, Y. (2024). Prompting large language models for zero-shot essay scoring via multi-trait specialization. *arXiv*. https://doi.org/10.48550/arXiv.2404.04941

Yancey, K. P., LaFlair, G. T., Verardi, A. R., & Burstein, J. (2023). Rating short L2 essays on the CEFR scale with GPT-4. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, & T. Zesch (Eds.), *Proceedings of the 18th Workshop on*

Innovative Use of NLP for Building Educational Applications (BEA 2023) (pp. 576-584). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.49

Jinmin Chung & Sungyeun Kim

Evaluating Validity of the Automatic Item Generation using Differential Item Functioning from a Multidimensional IRT

Abstract: Automatic item generation (AIG) is a new generation of model-based assessment methods in which computer algorithms generate items along with estimates of psychometric parameters. This allows for the creation of not only test items but also conceptually and statistically equivalent versions of the test. Developing a variety of items in the K-12 space requires a deep understanding of both the subject matter and the purpose of the assessment, making it time-consuming and expensive to create. AIG has gained a lot and is rapidly being adopted in K-12 because computerized systems can generate large numbers of high-quality items quickly and efficiently. However, while the new technology makes testing easier to administer to students and teachers, it raises questions about test security and validity (Patel, 2021). Validation is a key process that ensures the reliability and validity of an assessment tool and provides a scientific basis for score interpretation (American Educational Research Association, 2018). Item development is essential to validity assessment because it documents the procedures and outcomes necessary to create high-quality test content (Gierl et al., 2022). Most knowledge tests in various K-12 content areas are based on multiple item types, not just multiple-choice items. The multiple-question format item model (MQ-FIM) used in this study refers to the inclusion of multiple item types in one item model, including multiple-choice, fill-in-the-blank, and true/false, which are the most commonly used item types in K-12 testing. By separating what is being asked from how it is being asked, MQ-FIM helps us gain a deeper understanding of the construct being modeled.

There are few studies with real data on the utility and validity of these AIGs, and psychometric evaluations, such as equivalence between different tests, remain unclear (Falcao et al., 2022; Gierl et al., 2022; Rafatbakhsh et al., 2020). Since incorporating cognitive models into test design and development is necessary to support validity arguments for test-based reasoning, it can be assumed that AIG incorporates validity evidence into its methods (Gierl et al., 2022; Leighton & Gierl, 2011). Thus, by obtaining evidence of validity, this study will take an important step toward supporting the wider use of AIGs, including MQ-FIM, in educational assessment.

In this study, we use a simulation study to pilot test a high school diagnostic mathematics assessment. The AIG test is compared to a traditional, fixed format test with and without the application of a MQ-FIM. The validity of the AIG test is examined by utilizing differential item functioning (DIF) from the perspective of two-dimensional multidimensional item response theory (MIRT).

Even if two groups have the same underlying two-dimensional distribution, DIF can still occur depending on the test type. DIF can be examined by comparing the validity sectors defined by Ackerman (1992) in the four formats mentioned above. Insights gained from item vectors and validity sectors can help psychometricians improve assessments, ensure validity, and understand the complex interactions of latent abilities.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, 29, 67-91. https://doi.org/10.1111/j.1745-3984.1992.tb00368.x

American Educational Research Association. (2018). *Standards for educational and psychological testing*. American Educational Research Association

Falcão, F., Costa, P., & Pêgo, J. M. (2022). Feasibility assurance: a review of automatic item generation in medical assessment. *Advances in Health Sciences Education*, 1-21.

Gierl, M., Swygert, K., Matovinovic, D., Kulesher, A., & Lai, H. (2022). Three sources of validation evidence needed to evaluate the quality of generated test items for medical licensure. *Teaching and Learning in Medicine*, 1-11.

Haladyna, T. M. (2012). Automatic item generation: A historical perspective. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 13-25). Nueva York: Routledge.

Kim, S. Y., Yeum, S. K., Chung, J. M., Yoon, K. I., & Park, S. E. (2023, March). *Multivariate Generalizability Theory for Reliability with Item Models: Industrial Mathematics Test Example*. Annual meeting of National Council on Measurement in Education (NCME), Chicago, IL.

Kim, S. Y., & Chung, J. M. (2023). Exploring the Reliability of an Assessment based on Automatic Item Generation Using the Multivariate Generalizability Theory, *Journal of Science Education, 47(2)*, 211-224.

Leighton, J. P., & Gierl, M. J. (2011). *The learning sciences in educational assessment: The role of cognitive models*. Cambridge University Press.

Patel, S. (2021). *Exploring the effect of occlusion on a computerized mental-rotation test: Implications for automatic item generation*. Louisiana Tech University.

Rafatbakhsh, E., Ahmadi, A., Moloodi, A., & Mehrpour, S. (2020). Development and validation of an Automatic Item Generation system for english idioms. *Educational Measurement: Issues and Practice, 40(2)*, 1-11. https://doi.org/10.1111/emip.12401

Mayank Kumar

Humanizing Automated Essay Scoring using multiple Large Language Models

Abstract: Evaluating thousands of candidate essays manually is a labor-intensive task. In order to overcome this many organizations have moved to Automated essay scoring (AES). AES systems typically rely on extensive pre-training with large essay datasets on similar topics. Key features used for training include word length, n-grams, grammar, sentence cohesion, and semantic word roles. The problem with this approach is firstly, it is difficult to obtain large amount of essays for training. Secondly, models trained on specific datasets may not generalize well to essays outside of that dataset, resulting in poorer performance on unseen data. Thirdly, training requires lot of time. Fourthly, and perhaps most significantly, it is unable to distinguish between an authentic essay and a fake one. For instance, if a candidate is required to write an essay on the topic "Disadvantages of Air pollution," but instead writes on "Disadvantages of Water pollution," the model will still award the candidate with points because, despite the essay's off-topic nature, the sentences are coherent, and there may be words like "pollution," "health," and "climate" that are shared by both "Disadvantages of Air pollution" and "Disadvantages of Water pollution" that could lead the system to believe that the essay is on "Disadvantages of Air pollution."
The method we propose makes use of some user defined rules for assigning score and multiple existing Large Language Models (LLMs) without any fine-tuning. Our model does not use Chat-GPT due to security and ethical concerns due to which many organizations may try avoid using it. Our AES system requires two prerequisites. First of all, it is necessary to have at least one model response that includes an essay summary. The second step is to break down our topic into three parts: the meaning of the topic, the benefits and drawbacks of the topic, and the conclusion. For each part, we then list a few keywords.
Some candidates write meaningless words to meet the word limit because they are aware that their essays will be graded by a machine. Our algorithm flags essays that contain large number of meaningless words. It then categorizes essays into three components using zero-shot models, then summarizes them for comparison with user-provided summaries. This is done because even though a candidate may write essay completely differently, the core idea of an essay should remain the same. A question-answering model then evaluates essays, awarding marks based on accurate answers. In our hybrid scoring model, we allocate marks as follows: 30% for spelling and grammar, 10% for sentiment analysis, 20% for matching central ideas, 20% for zero-shot topic categorization, and the remainder for question answering. To validate accuracy, we assessed 1000 essays with both human and machine graders, resulting in a correlation of around 0.8.

S3-2: Symposium - Integrating AI into Adaptive Testing

Chair: *Hyo Jeong Shin & Ji Hoon Ryoo*

Abstract: The impacts of artificial intelligence (AI) on the educational measurement field are diverse, and the area of adaptive testing is no exception. In addition to providing tailored testing experiences that maximize efficiency and accuracy, as is the case with most adaptive testing, harnessing AI for adaptive testing has the potential to further improve the validity, reliability, and efficiency of the assessments. The CLASS-Analytics platform has been developed to provide novel and scientific solutions in this field. It integrates advanced AI capabilities into adaptive testing with real-time data analytics to enhance both assessment and instruction.
In this symposium, presentations will cover a variety of applications of integrating AI into adaptive testing that the CLASS-Analytics platform has served. These presentations will illustrate several successful applications across multiple fields in educational assessments in Korea as well as for the MACAT company located in the UK.

Presenters:

Hyun Suk Ryoo, Sunyoung Bu, & Ji Hoon Ryoo

Introduction of CLASS-Analytics Platform

Abstract: The CLASS-Analytics platform is designed to streamline the administration of adaptive testing and automatic scoring. By leveraging advanced algorithms, CLASS-Analytics adjusts the difficulty of test questions in real-time based on the test taker's performance, ensuring a more personalized and accurate assessment. This approach not only optimizes the testing process but also provides immediate feedback and metrics, eliminating the delays typically associated with traditional data analysis methods. In this presentation, we introduce the technical aspects of the CLASS-Analytics platform. We characterize three features as follows: 1) an assessment-guided learning analytics platform, 2) the use of Item Response Theory (IRT) based models for accurate measurement, and 3) a fault-tolerant and cloud-based infrastructure with zero downtime. We also introduce current projects in instruction and assessment, focusing on how the AI-driven capabilities are at the core.

Yeongjin Jo, Rayeon Kim, & Ji Hoon Ryoo

Developing the Korean CAT of Reading Motivation

Abstract: In CLASS-Analytics, we developed a computerized adaptive testing (CAT) algorithm to assess reading motivation in Korean middle school students. The measure of reading motivation was developed using a six-point Likert scale, which constituted an item pool for CAT. The concept of the Reading Motivation CAT and its underlying algorithm were initially proposed by Davis et al. (2020). In light of the necessity for a Korean reading motivation CAT program, an investigation was conducted to ascertain which constructs would be most suitable for Korean students. Four factors were identified, namely personality, textual, social, and cultural.
This presentation will detail the development process of the reading motivation CAT program in CLASS-Analytics, demonstrating how the item pool was constructed and the post-hoc simulation results. In particular, construct validity was evaluated through confirmatory factor analysis. Following the exclusion of misfitting items, the bi-factor CAT process was fitted (Weiss and Gibbons, 2007). In this process, each specific factor proceeds similarly to a unidimensional CAT. During the estimation, we compared the MLE (maximum likelihood estimation) and EAP (expectation a priority) through the post-hoc simulation. The EAP showed a better recovery, and thus a reading motivation score with EAP was reported. The resulting algorithm was implemented within CLASS-analytics, and then a live CAT was conducted for middle school students.

Hyo Jeong Shin, Seewoo Li, & Salah Khalil

Design considerations for adaptive testing with automatic generated items

Abstract: Measuring cognitive and non-cognitive skills requires the development of valid and reliable tests, which is often a labor-intensive and time-consuming process. As generative artificial intelligence (AI) becomes more accessible, researchers and the public have harnessed the power of AI to automatically generate large numbers of items. However, more research needs to be done and more rigorous verification procedures need to be established to support the deployment of AIgenerated items on-the-fly. In this study, we present several design considerations for adaptive testing when tests are composed of both human-generated and AI-generated items. A specific example is the generation of test items for the MACAT critical thinking tests using the GPT-4 model developed by OpenAI. The GPT-4 model has demonstrated remarkable performance in generating human-like text, making it an ideal candidate for creating test items that can assess critical thinking skills. In this presentation, we illustrate how we explored the use of item response theory (IRT)-based measurement models to support the valid use of AI-generated items, so that both human-generated and AI-generated items could be loaded onto the CLASSAnalytics platform. This work is illustrated with items and designs from an operational digital critical thinking test. Our findings have implications for adaptive test designs that include both human-generated and AI-generated items for the broader application of AI in educational testing and assessment.

Seewoo Li, Hyun Suk Ryoo, Hyo Jeong Shin, & Ji Hoon Ryoo

The implementation and score reporting of the CAT

Abstract: The MACAT's Critical Thinking assessment reports seven abilities: overall Critical Thinking skill and six subscales, including Problem Solving, Analysis, Creative Thinking, Interpretation, Evaluation, and Reasoning (PACIER). To report these seven scores on the CLASS-Analytics platform through the adaptive test design, we use the item response theory (IRT)-based measurement models that are similar to those used in bi-factor models (see Wise & Gibbons, 2007), allowing for the calculation of all seven scores while minimizing test length.

In the presentation, we will introduce the adaptive testing algorithm that uses the expected a posteriori (EAP) for routing purposes during test administration and calculates the weighted likelihood estimates (WLE) to estimate the seven ability scores for final reporting purposes.

Finally, to facilitate user understanding of the score report, the scores are linearly transformed from the logit scale to a MACAT scale score with a mean of 100 and a standard deviation of 10. The score report is automatically generated on the CLASS-Analytics platform so that users can be informed of their performance in real time.

S3-3: Paper Session - AI Topics 1

Chair: *Eun Hye Ham*

Eun Hye Ham

Assessing GPT-based Automated Scoring Quality in Scenario-based Assessments: Exploring Different Prompting Strategies

Abstract: Assessing students' (Bewersdorff et al., 2023; Park et al., 2023; Moore et al., 2022; Lee, 2023). 21st-century skills, including communication, often involves analyzing their reactions to relevant situations in scenario-based assessments. However, the human scoring of constructed response questions in these assessments can be labor-intensive and time-consuming, limiting their broader application, particularly in adaptive testing contexts. Recent advancements in large language models (LLMs) like GPT, known for their high knowledge and inference capabilities, show promise for automating this process. Prior research has explored the potential for GPT to match human scoring consistency, which varied findings (Bewersdorff et al., 2023; Park et al., 2023; Moore et al., 2022; Lee, 2023).

This study aims to assess the comparability of GPT-based automated scoring with human scoring for scenario-based assessments do not necessarily require specific subject matter knowledge but instead rely on transferable problem-solving skills. It also seeks to determine how different integration strategies of scoring criteria and prompting techniques affect scoring quality. Unlike prior studies focusing on subject-specific responses, this research examines non-content-specific responses, thereby broadening the application scope of automated scoring.

Furthermore, the study addresses the gaps in research concerning which judgements benefit most from GPT-based scoring and which do not, using both (1) holistic and (2) analytic scoring approaches (with detailed dichotomous criteria). Specifically, this study illustrates 10 different strategies for integrating scoring criteria and prompting strategies, which depend on the combinations of scoring criterion types – (1) holistic or (2) analytic scoring criteria, with three prompting techniques (1) zero-shot (Kojima et al., 2022), (2) few-shot with chain-of-thought (Wei et al., 2022), and (3) fine-tuned (Zelikman et al., 2022), as outlined in Table 1.

The analysis included responses from 478 middle school students to a scenario-based task assessing their democratic communication skills, scored by experts and compared across GPT-generated scores. This study utilized the GPT Python API (OpenAI, 2023) for scoring. The GPT-generated scores were compared with human-scoring scores with respect to total and item-level and GPT-scoring data, using TAM package (Robitzsch et al., 2023) and sirt package (Robitzsch et al., 2024) in R.

Results indicated that the few-shot with chain-of-thought prompting strategies, when combining with an analytic scoring checklist, aligned most closely with human scoring, achieving a .72 correlation, as shown in Table 1. This approach also closely matched human scoring in terms of item-level score and item difficulty. However, integrating GPT-4 scoring with human scoring data revealed significant discrepancies in both item difficulty parameters and item-fit, rater-fit indices in the man-faceted Rater model,

indicating limitations in the comparability of GPT-4 and human scoring outcomes. These findings will be detailed in the final paper and presentation. This study highlights the potential of integrating GPT with 21st century skill assessments to enhance the efficiency and scalability. While the results are promising, the discrepancies between human and GPT-generated scores call for further refinement of integration strategies. Future research should focus on optimizing these strategies and determining the conditions under which GPT can most effectively complement human judgment. Additionally, considering the inherent noise in human scoring, refining psychometric procedures to evaluate collaborative scoring between humans and LLMs is essential.

Reference

Bewersdorff, A., Seßler, K., Baur, A., Kasneci, E., & Nerdel, C. (2023). Assessing Student Errors in Experimentation Using Artificial Intelligence and Large Language Models: A Comparative Study with Human Raters. *arXiv preprint arXiv:2308.06088*.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large Language Models are Zero-Shot Reasoners. ArXiv. /abs/2205.11916 *Research, 61*(4). 299-332. http://dx.doi.org/10.30916/KERA.61.4.299.

Lee, G. G., Latif, E., Wu, X., Liu, N., & Zhai, X. (2024). Applying large language models and chain-of-thought for automatic scoring. Computers and Education: Artificial Intelligence, 6. https://doi.org/10.1016/j.caeai.2024.100213

Moore, S., Nguyen, H. A., Bier, N., Domadia, T., & Stamper, J. (2022). Assessing the quality of student-generated short answer questions using GPT-3. Paper presented at the 17th European Conference on Technology Enhanced Learning, Toulous, France.

OpenAI. (2023). *GPT-4 Technical Report*. http://arxiv.org/pdf/2303.08774v3

Park, S., Lee, B., Ham, E. H., Lee, Y., & Lee, S. (2023). Exploring the Possibility of Science-Inquiry Competence Assessment by ChatGPT-4: Comparisons with Human Evaluators. K*orean Journal of Educational Research, 61*(4). 299-332. http://dx.doi.org/10.30916/KERA.61.4.299

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. ArXiv. /abs/2201.11903

Tables

Table 1. Score Distributions and Correlations with Human-rated Scores by Variations in Integrating Scoring Criteria and Prompting Techniques

		Mean	Std. Dev	Pearson Correlation	Spearman Rank Correlation
Human scoring		4.55	3.18	-	-
GPT-scoring					

		Mean	Std. Dev	Pearson Correlation	Spearman Rank Correlation
Holistic Scoring	Zero-shot	9.11	2.87	0.49	0.45
	Zero-shot with CoT	10.72	2.98	0.56	0.65
	Few-shot	3.49	3.61	0.60	0.64
	Few-shot with CoT	6.41	3.80	0.68	0.68
	Fine-tuing	5.96	3.25	0.49	0.48
Analytic Scoring	Zero-shot	1.07	4.58	0.56	0.67
	Zero-shot with CoT	5.04	3.22	0.68	0.67
	Few-shot	3.34	3.98	0.58	0.68
	Few-shot with CoT	4.58	3.66	0.72	0.74
	Fine-tuning	3.32	2.32	0.25	0.24

Mikyoung Yim

Comparative study of ChatGPT and examinee performance on KMLE

Abstract: It was reported that ChatGPT passed the medical licensing exam in the United States. In one study, the scores on the 1st and 2nd stage tests exceeded the passing score for the example questions.
Accordingly, this study aims to evaluate the results of the Korean Medical Licensing Examination (KMLE). We would like to calculate the correct answer of ChatGPT for the KMLE administered in 2023 and compare this with the test taker's correct answer rate.
As a result, ChatGPT's correct response rate is compared with the average correct response rate of test takers, and this is also compared by subjects, specialties, and knowledge levels. Additionally, correlation with test taker responses is also evaluated.
When ChatGPT was launched, there were doubts about the accuracy of the information. We study how version 4.0 responds to the Korean Medical Licensing Examination in 2024. By specifically analyzing what kind of achievements are shown in which areas, we can get implications on how to use it in predicting test taker scores, creating and reviewing questions.

Xiangen Hu

Socratic Playground for Learning (SPL): A generative AI enabled platform combines assessment with tutoring

Abstract: We present the 'Socratic Playground for Learning,' a pioneering platform that harnesses the power of generative AI to seamlessly integrate assessment with personalized tutoring. The Socratic Playground leverages cutting-edge AI technology to create an adaptive learning environment where students engage in Socratic dialogue, enhancing critical thinking and problem-solving skills. The key feature of the Socratic Playground is its ability to assess students' understanding and ask deep reasoning questions tailored to their specific needs and levels of comprehension.

The platform incorporates several essential assessment elements, ensuring a comprehensive evaluation of student learning. Diagnostic questioning identifies initial knowledge levels, allowing the system to tailor the learning experience from the start. Formative assessment through continuous dialogue provides immediate feedback, enabling real-time adjustments to instruction. Critical thinking evaluation analyzes students' reasoning abilities, promoting intellectual rigor and deeper understanding.

Reflection and metacognition are emphasized, encouraging students to evaluate their cognitive processes and develop self-awareness. Application and transfer questions challenge learners to apply concepts to new situations, assessing their ability to transfer knowledge across contexts.

Conceptual understanding is a focal point, ensuring that students grasp underlying principles rather than relying on rote memorization.

Dialogue and discussion facilitate communication skills, as students articulate their thoughts and engage in intellectual discourse. Peer assessment promotes collaborative learning, allowing students to evaluate each other's responses and learn from diverse perspectives. Summative assessment synthesizes the discussion, gauging overall mastery and understanding of the topic.

This talk will present the initial implementation of the Socratic Playground and the design principles that guided its development. We will delve into the practical aspects of creating an adaptive, AI-driven learning environment that embodies the Socratic method. An efficacy study is currently being conducted to evaluate the platform's effectiveness, and we will share initial data on its impact on student learning outcomes.

Through these multifaceted assessment strategies, the Socratic Playground not only evaluates student performance but also offers real-time, targeted feedback and guidance, fostering a deeper understanding of the subject matter. This innovative approach transforms traditional educational paradigms, delivering a holistic, student-centered learning experience that empowers learners to achieve their full potential.

We will explore how the Socratic Playground leverages the power of generative AI to create a dynamic and engaging learning environment. The initial findings from our efficacy study will be discussed, highlighting the platform's potential to revolutionize education by combining assessment with personalized tutoring. This presentation aims to provide insights into the future of education, where adaptive learning and continuous assessment drive student success and mastery.

S3-4: Invited Symposium - Remembering Theo Eggen: A Symposium to Honor His Intelletual Legacy

Chair: *Wim J. van der Linden*

Presenters:

1. Bernard Veldkamp
   Improving the quality of examination, both in the Netherlands and abroad.
2. Nate Thomson
   Classification CAT: Do we even need scores?
3. Maaike van Groen
   The Dutch journey of the SPRT
4. Angela Verschoor
   Difficulty and exposure control in a mathematics test for teacher colleges

S4-0: Symposium- Current CAT Research at the University of Minnesota

Chair: *David Weiss*

Abstract: This symposium will highlight some of the current CAT research underway in the Department of Psychology at the University of Minnesota. It consists of six papers to be presented by five current graduate students and one recent Ph.D. from the CAT/IRT Lab. Each paper presents research that has been independently developed and pursued by its presenter, but all members of the Lab have assisted with the research design and analysis of each paper.

The first two papers are based on the adaptive measurement of change (AMC) method that has been one focus of the Lab for about a decade. The first paper presents results designed to evaluate post-hoc methods for identifying where in a series of three or more AMC CAT θ estimates significant change has occurred, when the overall AMC significance test has identified psychometrically significant change. The second paper applies the AMC method to the identification of significant high or low θ estimates for a multi-scale profile of scores for a single examinee, as would result from a person's profile of a multiscale CAT version of the Big Five personality inventory.

The third and fourth papers address, from different perspectives, an important and not previously adequately solved problem in variable-length CAT—that of terminating a CAT using a specified standard error of measurement stopping rule with item banks that do not provide equiprecise measurements across the θ scale. Paper 3 reports on the development and evaluation of a method of stochastic curtailment while the fourth addresses the same problem by proposing and evaluating a procedure based on a conditional standard of measurement perspective.

Paper 5 address an issue that will likely arise more frequently in the future now that multidimensional IRT can be implemented—the question of whether, with an instrument that has a between-items multidimensional structure, at what degree of dimensional intercorrelations a multidimensional CAT provides increases in efficiency and/or precision over treating the structure as a set of unidimensional CATs. The sixth and final paper presents data that demonstrates that predictive validities can be increased using the IRT-based standard errors of measurement as a moderator variable when predicting external criteria from CAT θ estimates.

Presenters:

Raj Wahlquist & David J. Weiss

Contrast Pairwise Comparisons in Adaptive Measurement of Individual Change

Abstract: For many years, testing to measure change or growth in education and psychology has been based on statistical procedures focused on measuring group differences rather than individual change. However, through computerized adaptive testing, we can now reframe this as a psychometric challenge rather than a statistical issue. By using item response theory (IRT) processes, computerized adaptive testing (CAT) allows us to estimate an individual's trait level in real time and select items that best match their estimated ability level (Weiss, 1985; Weiss & Şahin, 2024).

If a CAT is administered at multiple occasions, such as when monitoring educational progress or patient symptoms in hospitals, we can measure if significant change in the latent trait, $\theta\theta$, has occurred for the examinee over time. This procedure is known as adaptive measure of change (AMC), which analyzes the differences in the estimated trait levels of an individual at two or more occasions (Kim-Kang & Weiss, 2008). Currently, the best way to implement AMC over more than two measurement occasions, is through a likelihood ratio test comparing the log likelihood across two or more occasions. Specifically, Phadke (2017) found that the likelihood ratio test had the most optimal balance of appropriate Type 1 error and power over other procedures. However, while the likelihood ratio test can tell if a significant change has occurred over multiple occasions, there is currently not a standard post-hoc analysis method to determine where in a series of AMC CATs significant change has occurred and where it has not.

This monte-carlo simulation study will identify the best method for conducting pair-wise comparisons, as well as comparisons across larger subsets of estimated $\theta\theta$s, within the AMC framework. 1,000 simulees will be generated, at equal intervals, across the $\theta\theta$ continuum from -3 to 3 with a distance of 0.5 between intervals. Four different levels of change will be simulated: no change, small change, medium change, and large change (m = {0, 0.5, 1, 1.5}, respectively). Two item banks will be used for the CATs: an approximately uniformly distributed bank and a peaked item bank, both consisting of 500 items. Test length will vary between 15, 30, and 60 items. Lastly, the performance of three different methods for calculating pairwise comparisons will be evaluated by Type 1 error, power, and positive predictive value. The three AMC methods to be evaluated are the likelihood ratio test and the score test methods, as recommended by Wang & Weiss (2018), and a psychometrically based Z-test.

This work has the potential to be extremely important for any field that needs to measure individual change over many occasions and determine when psychometrically significant change has occurred

for an individual. For example, clinical psychologists who would be interested in measuring mental health changes for their patients. If clinicians can accurately identify when significant change has occurred over time, then they can better tailor treatments for individual patients. This, in turn, would lead to more efficient patient treatment and greater overall success in treating mental health.

References

Kim-Kang, G., & Weiss, D. J. (2008). Adaptive measurement of individual change. *Zeitschrift für Psychologie/Journal of Psychology, 216*(1), 49-58.

Phadke, C. (2017). *Measuring intra-individual change at two or more occasions with hypothesis testing methods* (Unpublished doctoral dissertation). University of Minnesota.

Wang, C., & Weiss, D. J. (2018). Multivariate hypothesis testing methods for evaluating significant individual change. *Applied Psychological Measurement*, 42(3), 221-239.

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology*, 53(6), 774.

Weiss, D. J. & Şahin, A. (2024). *Computerized adaptive testing: From concept to implementation*. Guilford Press.

Matthew A. Snodgress & David J. Weiss

Adaptive Profile Difference Analysis with Applications to Personality Assessment

Abstract: One unexplored application of AMC is to the detection of intra-individual, psychometrically significant differences among multiple traits for measurements obtained at a single occasion for a single examinee. Rather than administering one CAT on each occasion, a CAT would instead be administered for each trait with AMC's hypothesis tests applied to detect significant differences among traits using IRT-based trait estimates. For example, if an individual's score on a measure of extraversion differs significantly from the same individual's score on a measure of agreeableness, knowing whether these two personality traits differ significantly could provide useful information about an individual's personality tendencies. Extending the concept to all Big Five personality traits (Costa & McCrae, 1992), understanding how such traits differ within a single person could be used to tailor job training or educational interventions. Extended beyond personality to a comprehensive educational assessment, understanding whether scores in subjects like science and mathematics differ from scores in reading could be useful when prioritizing subjects that should receive greater attention for learning for each student. More generally, this procedure, denoted adaptive profile difference analysis, could provide greater granularity to multiscale assessments, improving interpretation of such assessments.

In this study, existing AMC omnibus hypotheses tests will be applied to the detection of differences on θ estimates across multiple traits for a single examinee. Simulations will be conducted using synthetic data based on two real, openly available personality datasets including: (1) a 50-item assessment of the Big Five based on the International Personality Item Pool (Goldberg, 1999); and (2) a 163-item assessment of Cattell's 16 Personality Factors also based on the International Personality Item Pool (Goldberg, 1999). The simulation study will address the following seven questions:

1. Do AMC omnibus tests detect psychometrically significant differences accurately?
2. Do tests have enough power to detect psychometrically significant differences?
3. How does performance change as the magnitude of differences changes?
4. How does performance change as the number of significant differences changes?
5. How do differences in scale information functions impact performance?
6. How does performance change for different numbers of scales being compared?
7. How does performance change for different omnibus tests?

Two outcome measures will include the false positive rate (i.e., the proportion of identified significant differences when there are no true

differences) and the true positive rate (i.e., the proportion of significant differences that are detected when there are true significant differences). Results will be discussed in terms of applications for educational and personality assessment. Future directions will also be discussed for further development of adaptive profile difference analysis.

References

Costa, P. T., & McCrae, R. R. (1992). NEO PI-R: Professional Manual: Revised NEO PI-R and NEO-FFI. Florida: Psychological Assessment Resources, Inc.

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychology in Europe, 7(*1), 7-28.

Ming Him Tai & David J. Weiss

Stochastic Curtailment: A New Approach to Improve Efficiency of Variable-Length CATs

Abstract: Stochastic curtailment (SC) is a statistical procedure that was originally developed to enhance the efficiency of clinical trials. It has been applied to psychological testing, but to sequential mastery testing only (Finkelman, 2008, 2010). This study adapted the method to detect low-precision examinees (i.e., examinees whose final standard error of measurement (FSEM) at the end of a full-length test could not reach the pre-specified SEM termination level) in measurement computerized adaptive tests (CATs). This can occur because (1) real CAT item banks generally fail to meet the desired uniform information function required for equiprecise measurement and/or (2) examinees do not always respond in accordance with a specified IRT model, which results in increased SEMs for those examinees. Using central limit approximations, the study developed a method to estimate the distribution of test information at maximum test length and the corresponding FSEM. The study also developed a hypothesis testing procedure to implement SC. Using monte-carlo simulations, the study found that (1) the FSEM estimation procedure performed well in the middle range of $\theta\theta$ values but less so at extreme $\theta\theta$ values; (2) the SC procedure had good predictive accuracy, with excellent performance on positive predictive values and good performance on true positive rates and false positive rates; (3) the reduction in test length was substantial. Overall, the study showed that SC is a promising procedure to identify low-precision examinees and enhance efficiency in measurement CATs. A guide on implementing SC is provided.

References

Finkelman, M. (2008). On using stochastic curtailment to shorten the SPRT in sequential mastery testing. *Journal of Educational and Behavioral Statistics, 33(4).*

Finkelman, M. (2010). Variations on stochastic curtailment in sequential mastery testing. *Applied Psychological Measurement 34(1).*

Joseph N. DeWeese & David J. Weiss

A Conditional Standard Error of Measurement Stopping Rule for Variable-Length CAT

Abstract: As compared to fixed-length CAT, variable-length CAT provides greater flexibility in balancing measurement precision and test length. A common termination criterion (also known as a stopping rule) for variable-length CAT is standard-error based termination, where an examinee's CAT is terminated once their standard error of measurement (SEM) falls below some prespecified cutoff value (Weiss & Kingsbury, 1984). In theory, the standard error stopping rule allows a CAT to achieve equal measurement precision across the latent trait (θ) range; however, this is only achievable with an item bank that provide essentially uniform information across the θ range. In practice, item banks tend to provide high information near the center of the θ distribution, or elsewhere, but lack information in other portions of the θ scale (e.g., Weiss & Sahin, 2024, Figure 15-1). As a result, many examinees can never achieve the desired SEM and are administered a test with an arbitrary maximum length.

Various solutions have been proposed to address inadequacies of the SEM rule, such as combining the SEM rule with secondary stopping rules intended to take effect in low-information regions of the item bank (Babcock & Weiss, 2012; Wang et al., 2019). We propose a novel solution: allowing the SEM cutoff value to vary as a function of the estimated latent trait () value. Using nonparametric quantile regression (Koenker, 2005) trained on simulated or real data, specific quantiles (such as the median) of the SEM for a test of a given test length can be predicted from . Once the model is trained on a set of θ and SEMs, it is simple and computationally efficient to apply it within a CAT by plugging in an examinee's after each item and predicting a specified quantile of the expected SEM at maximum test length. If we desire 50% of examinees to have a test that stops prior to the maximum test length, the quantile we choose would be the median. If we desire shorter tests or greater measurement precision, different quantiles can be chosen accordingly. In this way, the SEM cutoff value is intelligently chosen to reflect achievable levels of measurement precision throughout the entire θ range.

This study provides an initial examination of the efficacy of this new conditional SEM stopping rule in simulated data and compares it to existing alternative stopping rules, such as the original SEM rule, SEM plus change in θ rule, and the predicted standard error of measurement rule (PSER; Choi et al., 2011). The stopping rules will be compared across two item banks (a peaked realistic bank and a flatter approximately ideal bank), two maximum test lengths (20 items and 30 items), and two IRT models (3PL and GRM; Birnbaum, 1968; Samejima, 1969). The performance of each method will be evaluated based on the average test length, proportion of examinees reaching maximum test length, average SEM, bias of , and root mean square

error of ,. Future work will examine the stopping rules under a broader set of conditions and with real data.

References

Birnbaum, A. (1986). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Addison-Wesley.

Choi, S. W., Grady, M. W., & Dodd, B. G. (2011). A New Stopping Rule for Computerized Adaptive Testing. *Educational and Psychological Measurement, 71*(1), 37-53. https://doi.org/10.1177/0013164410387338

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press. https://doi.org/10.1017/CBO9780511754098

Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph. No. 17.*

Wang, C., Weiss, D. J., & Shang, Z. (2019). Variable-Length Stopping Rules for Multidimensional Computerized Adaptive Testing. *Psychometrika, 84*(3), 749-771. https://doi.org/10.1007/s11336-018-9644-7

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*(4), 361-375. https://doi.org/10.1111/j.1745-3984.1984.tb01040.x

Weiss, D. J. & Sahin, A. (2024). *Computerized adaptive testing: From concept to implementation*. Guilford Press.

Robert Chapman & David J. Weiss

Influence of Multidimensionality on IRT Modeling and CAT Administration

Abstract: To a greater or lesser extent, multidimensionality exists in many of the constructs we measure, such as math and reading ability or symptoms of depression and anxiety. Because of this latent multidimensionality, researchers are often faced with a decision about whether to utilize advanced multidimensional-specific IRT modeling methods to measure their constructs of interest. The present work seeks to provide guidance to the measurement and CAT research community regarding when to use multidimensional versus unidimensional IRT models. Monte-carlo simulation methods are used to evaluate realistic conditions (e.g., sample size, scale intercorrelations).

The specific research question is: At what magnitude of scale intercorrelation and sample size does a "between-item" multidimensional IRT model and multidimensional CAT perform better than a unidimensional CAT, in terms of administration efficiency and recovery of person parameters?

The study will consist of two stages. In the first "modeling" or "calibration" stage, multidimensional and unidimensional models will be fit to simulated data based on a set of monte-carlo design parameters. These parameters include number of dimensions (3 or 5), scale intercorrelations (r ranging from 0.2 to 0.7), item bank information shape (peaked or flat information for each dimension), and sample size (N= 300, 500, 750, and 1,000). Simulated data will be generated based on a between-items multidimensional graded response model with 50 items per dimension, and the multidimensional graded response model will be fit using R statistical software and the MIRT package for model estimation. In the second "implementation" stage, the IRT item parameters developed in the first stage will be applied as either independent unidimensional CATs or combined as "between-items" multidimensional CATs to score responses from a new set of simulees. Both stages of research will be evaluated in terms of person parameter recovery as indexed by bias and RMSE, and the second stage will additionally be evaluated in terms of CAT efficiency (number of items administered) and score precision (SEM).

Understanding how early research choices in the initial calibration phase later influence the CAT administration phase will inform researcher decision-making when implementing multidimensional IRT models for the purpose of measuring individuals.

Jesus Delgado & David J. Weiss

Using CAT Standard Errors of Measurement to Enhance Predictive Validity

Abstract: Tests or assessments are commonly used instruments designed to capture an examinee's competencies, abilities, or other psychological or educational variables in a given domain. All tests, however, are not created equally, with some yielding greater precision than others. Fixed item tests (FITs), for example, typically constructed according to the rules of classical test theory (CTT), generally possess large levels of IRT information around central levels of ability, but rapidly diminishing information is available to individuals as their ability levels deviate from the center of the ability distribution. Contrary to this, computerized adaptive tests (CATs) tailor the items presented to each examinee, such that these items are most informative for the examinee in question, thus reducing the observed standard errors of measurement (SEM) that indexes the (im)precision of θ estimates. In CTT , criterion-related validity is predicted to be a function of reliability—increased reliability is assumed to result in increased validity—and reliability can be increased using the correction for attenuation. However, no such relationship has been identified within the IRT context, because measurement precision (low SEM) can be conditional on θ, thus complicating the potential relationship between measurement precision and validity.

This research examined the relationship between IRT θ estimate precision (SEM) and an external criterion. Using real data from CATs administered in a medical environment, the data show that in most cases, after accounting for range restriction (and enhancement), greater precision in θ estimates (i.e., lower SEMs) is related to higher predictive validity, evidenced through stronger Pearson correlations with external criteria. These results, demonstrated through classical moderator analysis as well as statistical moderation analysis, will be presented. The results have also been replicated in a series of educational assessments using FITs. In addition, a monte-carlo simulation evaluated the improvements in θ estimate precision resulting from CATs when compared to FITs. Results show that for a broad range of θ estimate distributions, the administration of CAT reliably resulted in more precise θ estimates when compared to those obtained from comparable FITs, thus suggesting higher validities for CATs. Overall, these findings support the long-standing assertion that CAT can offer more precise measurements than FITs, and that this improved measurement is also related to improved predictive validity. Given that virtually all psychological and educational measurements are expected to be related to external criteria (e.g., relating standardized testing scores to academic performance) it is prudent to consider these benefits of CATs over FITs when creating and administering tests or assessments.

S4-1: Paper Session - Item Banking

Chair: *Ren Jie*

Ren Jie, Bu Tingfei, & Liu Dailichen

Realization of Multistage Adaptive Testing in National Common Language Proficiency Test of China

Abstract: It has been more than 20years since China started her National Common Language Proficiency Testin ethnic regions(usually referred to as MHK).In each test, MHK uses the same paper and pencil test paper to examine all candidates. This test format is known to lead to size able measurement errors for candidates at both ends of the ability distribution curve, resulting in a decrease in the overall test reliability. In this talk, I describe our three-year-long effort on the development of a computerized multistage adaptive testing scheme for MHK(referred to as MHK-MST).The benefit of MHK-MST is significant: compared to its predecessor, MHK-MST has fewer adaptive points and faster data processing, which is expected to significantly improve the test's reliability and validity.
My talk is divided into three parts.
I will first describe our creation of a comprehensive item bank for MHK. The item bank is composed of items, IRT scaling statistical parameters such as difficulty and discrimination, and various descriptive parameters of the items. The descriptive parameters mainly include the ability level of the items, the theme, style, and function of the text, etc. Over the past three years, we have established a large-scale scientific item bank that contains more than 50 sets of test papers.
In the second part of my talk, I will describe the evaluation of MHK-MST. Based on the created item bank, five MST panels were constructed using 0-1 linear programming methods. Simulation tests were conducted using Ronreal examinee data, and the results were evaluated using indicators such as RMSE, ABS, Bias, and Standard Error. Based on the testing results, a study on the reliability and validity of MHK-MST was conducted. In terms of reliability, the study used Cronbach's α, test-retest reliability, test reliability in IRT, and IRT test information to analyze it; in terms of validity, according to modern validity theory, evidence based on content, internal structure, and relationships with other variables was collected to analyze it.
If time allows, I will lastly introduce our ongoing effort on the development of a homemade software package that carries out the execution and performance evaluation of MHK-MST.

Yuan Ge, George Liao, & Young Koung Kim

Comparison of Automated Test Assembly Approaches and Solvers

Abstract: Automated Test Assembly (ATA) is a sophisticated process integral to educational assessments, ensuring the strategic combination of test items to construct valid and reliable test forms. Our research delves into a comprehensive comparative analysis of various ATA methodologies and solvers, with a particular focus on their implementation in large-scale standardized assessments.

The main challenge in ATA lies in optimizing the assembly of test items while adhering to content specifications, statistical constraints, and other relevant factors. Our study will explore different ATA methodologies, including but not limited to Integer Programming and Linear Programming, which both offer robust frameworks for tackling the multifaceted nature of test assembly. The exploration of ATA methodologies will reveal their strengths and limitations in the context of educational assessments. Our research will dissect these methodologies to discern their suitability for various assessment scenarios. Alongside these methodologies, we will conduct a thorough assessment of solvers such as GLPK, lpSolve, SYMPHONY, and Gurobi, which are integral to the computational process of ATA. The criteria for evaluation will center on efficiency and accuracy, which are essential in identifying the most compatible methods and solvers for ATA. Specifically, our study will consider factors such as computational speed, user-friendliness, and the capacity to process diverse datasets. These solvers, each with unique attributes, will be tested against real-world ATA contexts to gauge their performance efficacy.

The significance of our research extends beyond the theoretical framework, offering actionable insights into the applications and ramifications of test assembly practices within assessment companies and organizations, who are tasked with the responsibility of ensuring that educational assessments are not only fair and equitable but also reflective of the diverse abilities and knowledge of test-takers. By exploring different ATA methodologies and solvers, our study aims to amplify the effectiveness of test assembly processes, thereby enhancing the overall quality of educational assessments. Educators, test developers, and policymakers stand to gain from the evidence-based recommendations that will emerge from this research. The systematic comparison of ATA approaches and solvers will pave the way for optimized test assembly processes. Moreover, we will address the practical issues and possible solutions in ATA for large-scale assessments. Our research will also consider the implications of ATA in practical test assembly for assessments featuring linear testlets. Our study will incorporate findings from the domain of computer-based testing, which have become increasingly relevant in the digital age of education.

In conclusion, our research plans to make a significant contribution to the field of educational assessment. By providing a thorough comparison of ATA methodologies and solvers, we will assist assessment agencies in making informed decisions that will ultimately benefit the educational community at large. The findings of our study will not only support the development of high-quality assessments but also cultivate a more nuanced comprehension of the complexities inherent in the test assembly process. 1The College Board

Rae Yeong Kim, Insub Shin, Sung Sik Lee, & Yun Joo Yoo

Automated assembly for computerized multistage testing using deep knowledge tracing

Abstract: Computerized multistage testing (MST) and computerized adaptive testing (CAT) are developed as alternatives to conventional paper-and-pencil tests. In MST, the test is divided into several stages and examinees are offered a set of items, called module, at each testing stage. In contrast, CAT presents a single adaptive item to the examinee at any given time. While CAT can maximize testing efficiency due to its high adaptivity, MST offers several practical advantages over CAT as follows. First, except for some "on-the-fly" approaches, MST test forms are usually assembled before test administration, which allows test developers to monitor and control test quality. Second, examinees can review and adjust their responses within each module, which can reduce their stress and anxiety. Third, the structure of the response data matrix obtained from MST is simpler than that from CAT, enabling traditional statistical analysis, such as differential item functioning. Most MSTs are constructed and administered based on item response theory (IRT). IRT aids in the construction of MST by providing intuitive interpretations of item characteristics and by placing examinee's ability status and item difficulty on the same scale. However, IRT relies on assumption of unidimensionality, which constrain the representation of examinees' ability status on a single-dimensional value. Examinees' latent status can exist in a more complex structure, and item responses may also be determined by intricate interactions between the examinee's state and the item characteristics. Deep knowledge Tracing (DKT) models can address these limitations by capturing the complex structure between examinees' latent states and item characteristics through deep model architecture. Therefore, we aims to propose automated assembly methods and routing rules for MST using DKT models. First, a DKT model is trained based on examinees' response data to the items in the item bank. Then, using the trained DKT model, we propose automated module assembly methods that fit the predetermined multistage structure based on heuristic algorithm. The objective function for module assembly is designed to either minimize the uncertainty of the response probability distribution or to maximize the expected response probability of the target set of items, depending on the purpose or environment of the test. Additionally, the same method is applied to the item response probability distributions for a given examinee to implement routing rules for the constructed multistage test. The performances of the proposed methods are explored using synthetic data generated from multi-dimensional IRT. The results confirm that the proposed method can construct multistage modules with different characteristics, with perspectives that reduce the uncertainty of estimation or enhance the examinees' performance on the target set of items. It is anticipated that, using the proposed methods, multistage tests can be constructed even when there exist complex relationship between examinees and items.

Haejin Kim

Enhancing Personalized Education through Item Response Theory: Addressing the Cold Start Problem with Imputed Values

Abstract: In the field of educational measurement, leveraging Item Response Theory (IRT) within intelligent tutoring systems (ITS) offers significant potential for personalized education. IRT, typically used for test-based assessments, can continuously evaluate a student's ability through their problem-solving processes during learning. However, a significant challenge arises with the cold start problem when new knowledge concepts emerge or at the beginning of new academic years, where no prior student information exists. This research explores methods to estimate item parameters even in the absence of response history, addressing the sparse matrix issue in personalized learning contexts.

In adaptive learning environments, not all students encounter the same problems, resulting in a highly sparse problem-student interaction matrix. Direct application of IRT in such settings can yield low accuracy in item parameter calibration due to incomplete information. This study hypothesizes that imputing values for problems not presented to the student can enhance the stability of parameter estimation. While traditional missing data handling methods consider missing values as intentionally or non-intentionally omitted or not-reached items, this research focuses on imputing values for genuinely unpresented items caused by the adaptive learning environment.

This study proposes an imputation method that mutually considers the correction proportion of the item and the correction proportion of the students. Compared to other imputation methods, such as single value imputation and multiple imputation, this mutual consideration method shows a high recall rate of the simulated original item parameters.

The findings demonstrate that using imputed values reduces the standard error of estimated item parameters compared to using sparse response matrices without imputation, thus improving the estimation's stability. This approach shows promise in mitigating the cold start problem. Future research could explore various imputation methods and refine strategies for stabilizing item parameter estimation in adaptive learning systems.

S4-2: Paper Session- CAT Applications 3

Chair: *Dorinde Korteling*

Dorinde Korteling, Selina Limmen, Marjolijn Ketelaar, Hedy A. van Oers, Manon A. T. Bloemen, Raoul H.H. Engelbert, Michiel A. J. Luijten, & Lotte Haverman

PROMIS® CATs Efficiency in Pediatric Physiotherapy Compared to Full Item Banks and Short Forms

Abstract:

Background: Patient-reported outcome measures (PROMs) capture patients' health perceptions through questionnaires, providing valuable information that enables healthcare providers to deliver personalized care. To effectively implement PROMs in pediatric physiotherapy, it is crucial to use broadly applicable measures that can accommodate the diverse functional capacities and diagnoses of children. Additionally, it is essential to address the lengthy application time of PROMs, which may lead to low response rates. Computerized Adaptive Tests (CATs) are promising in this field, as they minimize respondent burden while maximizing measurement precision by dynamically adjusting the difficulty of questions based on the respondent's previous answers. The Patient-Reported Outcomes Measurement Information System (PROMIS®) is a PROM system that allows for the use of CATs and is commonly used in (pediatric) healthcare.

Objective: To compare the reliability and efficiency of PROMIS® CATs against PROMIS® full-length item banks and short forms for Dutch children receiving pediatric physiotherapy, in order to explore the suitability of CATs in general, and specifically for use in pediatric physiotherapy.

Methods: 300 children and adolescents (8-17 years) treated by a pediatric physiotherapist within the last year, and their caregivers (of children aged 5-17 years), will complete three full pediatric or proxy PROMIS® item banks v2.0 (Pain Interference, Mobility, Upper Extremity) and five pediatric or proxy PROMIS® CATs (Anxiety, Depressive Symptoms, Fatigue, Peer Relationships, Sleep Disturbance). The distribution of scores obtained by patients will be examined to explore the diversity in abilities of the children treated by pediatric physiotherapists. Reliability of item banks, short forms (4 and 8 item version), and CATs was expressed as standard error of theta (SE(θ)) and compared to each other. To investigate the reliability of PROMIS® item banks on Pain Interference, Mobility and Upper Extremity, post-hoc CAT simulations will be performed. Post-hoc CAT simulations will be performed using maximum posterior weighted information selection criterion and the Expected A Posteriori estimator. The starting item will be the one providing the highest information at the study sample's mean. Stopping rules for the post-hoc CAT will be a maximum of eight items administered (the length of a short form) or a SE(θ) < 0.32, corresponding to a reliability of 0.90. The full item banks, short-forms and (post-hoc) CATs will be compared on efficiency.

Results: We will present the results of 300 participants at the conference. Currently, 232 children/adolescents and their caregivers have already

participated. Data collection and analyses are expected to be finished in summer of 2024.

Conclusions: This study is the first to investigate the reliability and efficiency of PROMIS® item banks, short forms and CATs for children and adolescents treated by a pediatric physiotherapist. We aim to demonstrate the suitability of PROMIS® CATs for use in pediatric physiotherapy as an efficient and broadly applicable alternative for full length item banks and short forms. The use of PROMIS® CATs in healthcare aids early and quick identification of potential healthcare-related problems in children receiving pediatric physiotherapeutic care and facilitates communication between patients, parents and physiotherapists.

Nathan Thompson

Adaptive Testing Simulations to Drive Item Pool Design for K-12 Assessment

Abstract:

Introduction: An essential step in the design of item pools for adaptive testing is to perform simulation studies based on information from the current version of the assessment. This study utilized a pool of items designed to assess three different content areas (writing, reading and mathematics) in 6th grade, aligned to the national curriculum in a Mexican state. The item pool was successfully reviewed by subject matter experts which made changes in the original version of the item pool. After that, a pilot study was implemented and the item pool was calibrated with Rasch model. The results of this were used for adaptive testing simulations to evaluate feasibility, item bank evaluation, and planning of future development.

Aim: The aim of this study was to evaluate how well CAT would work to improve the assessment.

Methodology:
A post-hoc CAT simulation study was conducted (n=398) with the following constraints:

1. Initial estimate of theta: 0
2. Item selection criteria: Maximum Fisher Information, in Rasch model it would be the closest b value for the current theta estimate.
3. Scoring method: EAP

We then evaluated variations in the termination criterion (0.35, 0.40, 0.45) and item bank size (current, double, double with new items in certain difficulty range).
The study was made with CATsim software (Weiss & Guyer, 2010).

Results: The results showed that the easiest items were used more frequently, as a result, the correlation between thetas was low 0.630 for the current pool. This evidence suggests that the item pool is just too difficult for the examinees, since the most difficult items were not used in the simulation process.

Discussion: It is important to note that in the item review process, the subject matter experts pointed out that the item pool would be easy for the examinees, but the psychometric analysis showed just the opposite. Next steps may consider developing more items to increase the item pool size, specifically easy and medium difficulty.

Reference

Weiss, D. J. & Guyer, R. (2010). *Manual for CATSim: Comprehensive simulation of computerized adaptive testing*. St. Paul MN: Assessment Systems Corporation.

Istiani, M. Psi, Wenny Chatarina, & S. Psi, Maryanto

Comparing item estimation methods in personality tests for driving licenses (SIM) in Indonesia

Abstract: Personality tests are a crucial component of the driving license examination (SIM) in Indonesia. Aiming to assess psychological characteristics of prospective drivers relevant to road safety in Indonesia. The validity and reliability of test are paramount for producing accurate and fair assessments. This study compares various item estimation methods within the framework of Item Response Theory (IRT) for personality tests used in SIM examinations in Indonesia. This study aims are to be compared which item estimation method is most appropriate, between the Partial Credit Model (PCM) and the Graded Response Model (GRM).

This study utilizes data from SIM exams involving many respondents (N= 824.723) from various region of Indonesia. The number of items for personality inventory are 60 and the dimensions are prosocial behaviour, emotion stability, self-control, and adjustment. Item format in Likert style with 5 choices from strongly disagree until strongly agree. Item parameter estimation is conducted using R software.

Using Akaike Information Criteria (AIC) to solve the problem of model selection. AIC is formulated for selecting the 'best estimate' model among several measurement models with quantities different parameters, based on appropriate statistical criteria. The results obtained are that the 3PL (or Improvement) model has the lowest AIC index between the GRM (82182703) and PCM models (132972133). It means the GRM Improvement model has better suitability to the data compared to the PCM model. The second consideration is from ANOVA analysis shown the significant index ($p<0.05$). Based on the appropriate model, the number of items that are less suitable is 21 items out of 60 items.

Findings indicate that the 2PL and 3PL models provide more flexible and accurate estimations compared to the Rasch Model, particularly in identifying items with varying discrimination parameters. The 3PL model, which considers guessing parameters, also demonstrates better fit in the context of personality tests. Additionally, DIF analysis reveals some items functioning differently across different groups, highlighting the need for item revision to ensure fairness, but for the analysis is nor include the DIF yet. The result In conclusion, the use of IRT in GRM improvement models in item analysis of personality tests for SIM examinations in Indonesia is recommended due to their higher flexibility and accuracy. These findings are expected to contribute to the development of more effective and fair personality tests, enhancing road safety through more precise selection of prospective drivers.

S4-3: Paper Session - AI Topics 2

Chair: *Young Koung Kim*

Young Koung Kim & Tim Moses

Modeling AI Generated Essay Detection using ChatGPT Generated Essays

Abstract: Artificial Intelligence (AI) tools that generate text, such as ChatGPT, have gained rapid popularity in education. This popularity raises concerns about academic integrity and AI-assisted cheating. In writing assessments, AI tools threaten the accurate evaluation of students' true writing abilities. There is a need for methods to distinguish text that is generated by AI tools from text written by humans. Although many organizations provide AI detection tips and tools (e.g. Turnitin, AI Classifier by OpenAI etc.), their detection accuracy remains unknown.

The purpose of this study is to understand patterns of text generated by ChatGPT and examine factors that affect detection accuracy. Responses to two essay prompts from real world data generated by ChatGPT and written by students are collected and combined into one dataset. Large language models such as Bidirectional Encoder Representations from Transformers model (BERT) for classification are used to identify ChatGPT generated text versus student generated text. Several machine learning models are also examined and the detection accuracies of the models are compared. Lastly, the impact of modifications in ChatGPT generated text on the detection accuracy is examined.

This study uses two prompts from an Automated Student Assessment Prize (ASAP) dataset from the 2012 Kaggle competition1.) The gpt-3.5-turbo version of ChatGPT (ChatGPT3.5) generated 1,376 essays and 1,836 essays for Prompt 1 and Prompt 2 from the ASAP data, respectively. Since the default values for ChatGPT3.5's textual randomness levels were used for initial results, all the essays generated by ChatGPT3.5 have similar structures. In reality, students are more likely to use parts of the AI generated text rather than the entire text. Therefore, four conditions are examined depending on the modification of the essays - No modifications, extract the First three sentences, extract the Mid three sentences, extract the Last three sentences, and Mix the ChatGPT generated texts with the human generated texts2.) For the study analyses, the ChatGPT essays were combined with human generated essays and detection rates are computed for all study conditions.

As AI generated texts become more popular, identifying texts generated by AI tools have important implications for educational exams, cheating detection, and accurate assessments of writing ability. Preliminary results show that AI generated texts can be accurately achieved using simple machine learning models as long as all text for prediction are from AI tools. Once the sources of the texts are mixed, which is closer to reality, the predicted detection accuracy drops. Large language models like BERT may perform better than machine learning models in these cases. The final paper will include additional analyses of cases where ChatGPT text is generated with complicated instructions (e.g. "Write an essay like ones generated by human").

1.) The ASAP data contains 8 essay prompts (and their responses) consisting of persuasive/narrative/expository prompts and source-dependent response prompt (https://www.kaggle.com/competitions/asap-aes/overview).
2.) Specifically, the first three sentences from human essays, the middle six sentences from ChatGPT essays and last three sentences from human essays were combined for the Mix condition.

Ryan Lerch & Young Koung Kim

Modeling for Detecting Essays with Invalid Responses in Automated Essay Scoring

Abstract: The purpose of this study is to develop models that can accurately identify invalid responses in essays written by students. Invalid responses in essay writing pose a significant challenge to the effectiveness of automated essay scoring systems. These responses include, but are not limited to, off-topic responses that are unrelated to the essay prompts, and 'adversarial' responses that simply copy or repeat the prompts.

To address this issue, our study explores various Natural Language Processing (NLP) features and models. In our exploration of models, we will examine large language models, including Bidirectional Encoder Representations from Transformers (BERT) and variants of BERT models, due to their proven effectiveness in understanding the context of text. We also investigate traditional machine learning models that classify invalid responses, and we compare classification accuracies across all models.

We utilize the ASAP (Automated Student Assessment Prize) data as real data with 'valid' essay responses to various prompts. For simulations, we employ the GPT-4o tool to generate synthetic 'invalid' responses to the selected prompts. The extent to which essays are off-topic or invalid varies across the simulated essays. Upon analyzing these variations, we categorize the invalid responses into the following five types: (1) Fully-formed essays that simply restate and rephrase the prompt and make no original contribution; (2) Essays on a topic completely unrelated to the original prompt; (3) Eloquent but nonsensical essays that nevertheless reference keywords relevant to the original prompt; (4) Essays on a topic mostly unrelated to the original prompt, but with frequent references to relevant keywords; (5) Essays on a closely-related, but distinct topic to the original prompt.

These five response categories were chosen either because they represent the most common fully-formed invalid response types seen on large-scale assessments, or they are the types of invalid responses in greatest need of further study.

After generating the synthetic invalid responses, we will append them to the observed responses and then use the aforementioned models to classify responses as either valid or invalid. Our goal is not necessarily to score the essays; rather, we intend to use NLP models to identify essays that should and should not be scored.

This project is a work in progress and will be completed by the end of July. As proof of concept, we have reviewed the types of invalid responses commonly found in large-scale
assessments, examined the ASAP data, and shown that GPT-4o can generate adequate invalid responses for all five categories. GPT-4o can also generate responses of varying quality. We can therefore see if typographical and grammatical errors also influence the flagging of invalid responses.

Through this study, we aim to contribute to the improvement of automated essay scoring systems by enhancing their ability to accurately identify and handle invalid responses. This will be beneficial for ensuring scoring accuracy and validity.

Mfonobong Umobong, Godfrey Udo, Eme Joseph, Udeme Tommy, & Ukponobong Antia

Exploring Artificial Intelligence Technologies' Integration for Educational Assessment in Higher Educational Institutions in Nigeria

Abstract: The transformative potential of artificial intelligence (AI) in the educational sector is increasingly evident, particularly in the realm of assessment practices, personalized learning analytics and adaptive testing. Despite this, the scope of AI utilization in educational assessment within the higher education landscape of Nigeria remains under explored. This study seeks to bridge this knowledge gap by investigating the adoption, effectiveness, and challenges of AI technologies in academic assessment within this context.

Through a comprehensive mixed-methods approach involving quantitative surveys and qualitative interviews with 2000 faculty members across six universities in Nigeria's south-south geopolitical zone, the research delves into the awareness, usage patterns, perceived benefits, and obstacles of AI integration in educational assessment. Initial findings indicate a notable adoption rate of AI tools among lecturers yet reveal a diverse adoption landscape influenced by factors such as academic discipline, technological proficiency, and institutional support. The results highlight the potential benefits of AI in terms of improved efficiency and accuracy, personalized feedback, and reduced workload. The paper concludes with recommendations for policymakers and educational leaders to foster more effective and equitable integration of AI technologies in educational assessment. It also underscores the importance of government and institutional support as well as faculty development programs needed to bridge the gap between awareness and effective utilization of AI tools in assessment practices. Concerns regarding bias, ethical considerations, student privacy, over reliance on standardized data, and the need for adequate training emerged as key challenges.

S5-1: Paper Session - IRT Applications

Chair: *Ryan EK Man*

Ryan EK Man, Bao Sheng Loe, Eva K Fenwick, & Ecosse L Lamoureux

Method of Successive Dichotomization versus Andrich Rating Scale Model to Resolve Disordered Thresholds in Rasch Analysis: A Real-World Comparative Study

Abstract:

Background: Although simulated data have shown higher correlations between the method of successive dichotomization (MSD) model and "true" values in comparison to Andrich rating scale model (RSM) in the presence of disordered thresholds, ambiguity remains regarding the psychometric precision and agreement between real-world person measures and item calibrations derived from these two models. To assess the superiority of the MSD model relative to RSM in addressing disordered thresholds during Rasch analysis, this study compared the psychometric precision, mean difference (MD) and correlation between the person measures and item calibrations derived using these two models from two age-related macular degeneration (AMD) quality-of-life (QoL) item banks (IBs) that had disordered thresholds.

Methods: We recruited 261 patients with AMD (mean age 70.5±7.6 years; 40.6% female) from a tertiary eye clinic in Singapore to develop a series of QoL IBs, operationalized using computerized adaptive testing (CAT) to quantify the impact of the disease and related treatment from the patient's perspective. Using both MSD and RSM models, we evaluated the psychometric fit (i.e., person separation index [PSI] and person reliability [PR]) for two IBs that displayed disordered thresholds (Concerns [CN] and Lighting [LT]) and derived the associated person measures and item calibrations. MD between the person measures and item calibrations were determined using paired t-tests, while agreement was evaluated using intraclass correlation coefficient (ICC). Bland-Altman (BA) plots was also utilized to check for the presence of bias during agreement testing.

Results: The MSD model resulted in consistently greater psychometric precision indices (PSI and PR) across both QoL IBs compared to the RSM model (4.14 and 0.94 vs 2.67 and 0.88 for the CN IB; and 2.84 and 0.89 vs 1.93 and 0.79 for the LT IB). Higher person measure values were also observed for MSD vs RSM (MD: 1.30 and 1.24 logits for CN and LT, respectively; both P <0.05), although this trend was not present in item calibrations (MD: -0.02 [P=0.68] and 0.25 [P=0.03] logits for CN and LT, respectively). While ICC values were high for item calibrations (0.98 [0.97, 0.99] for CN and 0.97 [0.91, 0.99] for LT), the ICC for person measures were lower and more variable (0.85 [-0.09, 0.96] for CN and 0.84 [-0.08, 0.96] for LT). BA plots further showed clear evidence of proportional bias for the difference in person measures and item calibrations between the two models.

Conclusion: In our real-world dataset of IBs with disordered thresholds, the MSD model displayed superior psychometric precision relative to the RSM model. However, the agreement indices between the person measures and

item calibrations derived from these models were inconsistent with notable proportional bias observed. Future research should explore the implications of these disparities on the optimal number of items required to achieve satisfactory test precision in CAT simulations using these item calibrations, as well as their impact on the estimation of item exposure rates and the concordance of QoL scores between the CAT instrument and the full IBs.

Aikorkem Zhapparova & Nurym Shora

Implementing Computerized Adaptive Testing for Mathematics Monitoring Exam at Schools in Kazakhstan

Abstract: Ensuring an academic success and growth of students is the heart of mission at intellectual schools (IS) in Kazakhstan. In accordance with Academic Achievement Assessment Policy, a comprehensive monitoring framework was implemented for students in grades 7-12 , covering core subjects including Mathematics. The assessment in conducted two times a year in September and January, and designed to systematically track academic progress and improve the quality of education through continuous evaluation. The Mathematics monitoring test at IS is structured into five domains: Algebra, Geometry, Numbers, Statistics, and Mathematical Modeling. Each subtopic comprises 35 questions, with testing conducted over a period of five days. Alongside the assessment process, significant time is dedicated to analyzing student responses using Item Response Theory (IRT) and providing detailed feedback. Recognizing the need for a more time efficient and accurate approach to monitoring practices, there has been a transition towards Computerized Adaptive Testing (CAT). CAT adapts to students' abilities in real-time, departing from traditional linear testing methods. This work explores the case of shifting the 7th grade Mathematics monitoring exam towards CAT at IS in Kazakhstan.

Initially, existing test items were recalibrated using responses from tests between 2015 and 2021. Due to limitations in the number of items in the item bank, all domains were integrated into a single test. Subsequently, the items from each domain were analyzed for unidimensionality to ensure they collectively measure the construct of Mathematics. The correlation of parameters estimated for combined and separate domains exceeded 0.93, affirming the feasibility of combining items into a unified item bank. Following recalibration, ability scores for a sample were estimated using the Expected A-Posteriori (EAP) method and compared with previous scores. The correlation coefficients, surpassing 0.90, underscored strong consistency and reliability across assessments. A simulation study was then conducted using the mirtCAT package in R to optimize algorithm settings for the Mathematics monitoring test at. The simulation recommended displaying a maximum of 45 items to students, while ensuring balanced coverage across domains aligned with the curriculum.

The first CAT-format monitoring exam for 7th graders at IS in Kazakhstan was implemented in 2022 using the Concerto platform. This transition faced several challenges related to the item bank and software provisions , particularly with over 2500 students per grade. While there are ongoing challenges and improvements to address, we are already observing the efficiency and benefits of adaptive monitoring test. CAT significantly reduced test length, optimizing student assessment time, and facilitated immediate feedback delivery post-exam. The process of providing feedback to students on their results typically required up to two weeks before.The transition to CAT has already demonstrated notable efficiencies. Continued efforts will focus on overcoming challenges and maximizing the benefits of adaptive testing to further enhance educational outcomes at IS in Kazakhstan.

S5-2: Paper Session - Cognitive Diagnosis CAT 2

Chair: *Ivy P. Mejia*

Ivy P. Mejia, & Kevin Carl P. Santos

Mastery Profiles of Students in Explaining Phenomena Scientifically Under the G-DINA Model Framework

Abstract: Enhancing data-driven diagnostic assessment practices is imperative for classifying students' learning profiles effectively. This targeted approach ensures precise instruction and identification of students' mastery needs. Cognitive diagnosis models (CDMs) offer valuable insights into students' diagnostic profiles, crucial for mastering specific subjects. These models enable continuous assessment of essential competencies, even in scenarios like remote learning with limited connectivity. However, in the literature on the cognitive diagnosis model (CDM), the widely used disciplines applying this measurement model are the required skills in mathematics and reading and few in science education. In the current study, the model of cognition of students on explaining phenomena scientifically was explored and validated. The best fitting model and the diagnostic profiles of the students on the Use of Various Forms of Evidence to Explain Phenomena Scientifically (UVFEEPS) as a scientific inquiry skill were identified under the G-DINA model framework.

Ahoo ShokraieFard & Frank Back

Applying G-DINA model on a learning study about subtraction with negative numbers

Abstract: Diagnostic classification models offer the possibility of evaluating different aspects of knowledge through a quantitative analysis of test results (Maas et al., 2022). There is a great opportunity to study the implementation of DCM on real data in experimental educational designs (Paulsen & Valdivia, 2022). To apply DCM, the knowledge domain should be divided into the underlying construct, which character may be defined differently as the structure, abilities, strategies, or misconceptions of the knowledge (Shi et al., 2021). In a learning study, a kind of experimental classroom design based on variation theory, knowledge is achieved by discerning different aspects of the object of learning (Kullberg, 2010; Marton, 2015). Defining the critical aspects of the object of learning and developing test items that can relate to those aspects scaffold the basic requirements for applying DCM.
The object of learning in this study, subtraction with negative numbers, is defined by its critical aspects (CA):

CA1: The number system of integers
CA2: The twofold meaning of the minus sign
CA3: Subtraction can be interpreted as a difference
CA4: The commutative law does not apply in subtraction

The R-package G-DINA (Ma & de la Torre, 2020) is used as the main package, and the R-package CDM with G-DINA function (George et al., 2016) is used to compare and complete the analysis. Other methods, such as research team discussions based on think-aloud interviews, are mainly used to evaluate the validity of the analysis.

Having 128 test-takers, the analysis of critical aspects on the group level showed that CA1 and CA2 were discerned by over 80%, CA4 by 60%, and CA3 by 29% of participants. Almost 16% of test-takers discerned all four aspects. The analysis at the individual level demonstrated that pattern 1101 was the most common pattern (45% of test takers), meaning that CA1, CA2, and CA4 were discerned but not CA3. We also found out that a learner can be placed on a spectrum from how critical the aspect is to how well it is discerned; for example, for a participant with 1101, the probability of discerning CA3 was 30%, while for another participant, it was only 4%. To evaluate the accuracy of the analysis, we analyzed think-aloud interviews with the test-takers conducted directly after they took the written test. In those think-aloud interviews that we analyzed, we didn't find considerable contrasts with the statistical analysis generated by G-DINA.

The learning study design bridges the gap between learning and instruction (Marton, 2015). Employing DCM facilitates bridging assessment, learning, and instruction (Lee & Sawaki, 2009; Ren et al., 2021). The main obstacles that researchers face during the implementation of DCM are designing the test items, the Q-matrix specification, and the sample size (de la Torre, 2011; Lei & Li, 2016). In our case, when we were satisfied with the Q-matrix, we simulated data to increase the number of test-takers from 128 to 1000, which improved the model-fit criteria.

References

de la Torre, J. (2011). The Generalized DINA Model Framework. *Psychometrika, 76*(2), 179-199. https://doi.org/10.1007/s11336-011-9207-7

George, A. C., Robitzsch, A., Kiefer, T., Groß, J., & Ünlü, A. (2016). The R package CDM for cognitive diagnosis models. *Journal of statistical software, 74*(2), 1-24. https://doi.org/10.18637/jss.v074.i02

Kullberg, A. (2010). *What is taught and what is learned : professional insights gained and shared by teachers of mathematics* Diss. Göteborg : Göteborgs universitet, 2010].

Lee, Y.-W., & Sawaki, Y. (2009). Cognitive Diagnosis Approaches to Language Assessment: An Overview. *Language assessment quarterly, 6*(3), 172-189. https://doi.org/10.1080/15434300902985108

Lei, P.-W., & Li, H. (2016). Performance of Fit Indices in Choosing Correct Cognitive Diagnostic Models and Q-Matrices. *Applied psychological measurement, 40*(6), 405-417. https://doi.org/10.1177/0146621616647954

Ma, W., & de la Torre, J. (2020). GDINA: An R Package for Cognitive Diagnosis Modeling. *Journal of statistical software*, 93(14), 1-26. https://doi.org/10.18637/jss.v093.i14

Maas, L., Brinkhuis, M. J. S., Kester, L., & Wijngaards-de Meij, L. (2022). Diagnostic Classification Models for Actionable Feedback in Education: Effects of Sample Size and Assessment Length. *Frontiers in education (Lausanne), 7*. https://doi.org/10.3389/feduc.2022.802828

Marton, F. (2015). *Necessary conditions of learning*. London : Routledge.

Paulsen, J., & Valdivia, D. S. (2022). Examining cognitive diagnostic modeling in classroom assessment conditions. *The Journal of experimental education, 90*(4), 916-933. https://doi.org/10.1080/00220973.2021.1891008

Ren, H., Xu, N., Lin, Y., Zhang, S., & Yang, T. (2021). Remedial Teaching and Learning From a Cognitive Diagnostic Model Perspective: Taking the Data Distribution Characteristics as an Example. Frontiers in psychology, 12, 628607-628607. https://doi.org/10.3389/fpsyg.2021.628607

Shi, Q., Ma, W., Robitzsch, A., Sorrel, M. A., & Man, K. (2021). Cognitively Diagnostic Analysis Using the G-DINA Model in R. Psych, 3(4), 812-835. https://doi.org/10.3390/psych3040052

S5-3: Paper Session - Constraint Management in CAT

Chair: *Kylie Gorney*

Kylie Gorney & Mark D. Reckase

Using Multiple Maximum Exposure Rates in Computerized Adaptive Testing

Abstract: In computerized adaptive testing, item exposure control methods are often used to provide a more balanced usage of the item pool. Many of the most popular methods, including the restricted method (Revuelta & Ponsoda, 1998), use a single maximum exposure rate to limit the proportion of times that each item is administered. However, Barrada et al. (2009) showed that by using multiple maximum exposure rates, it is sometimes possible to obtain an even more balanced usage of the item pool at little to no cost in measurement accuracy. Therefore, in this paper, we develop an extension of the restricted method that involves the use of multiple maximum exposure rates. The idea behind the method is to (a) promote the selection of less popular items at the beginning of the test and (b) reserve the selection of more popular items for the end of the test. To satisfy both conditions, we apply a different maximum exposure rate at each position of the test. Then, an item is only allowed to be administered in a given position if its observed exposure rate is less than the maximum exposure rate that is applied at that position. A detailed simulation study reveals that the new method is able to (a) provide a more balanced usage of the item pool and (b) improve measurement accuracy. Taken together, these results are highly encouraging, as they reveal that it is possible to improve both item pool utilization and measurement accuracy simultaneously.

Chia-Wen Chen, Joe Waston, & Bryan Maddox.

Restricting item difficulty jumps to alleviate negative test experiences during computerised adaptive testing.

Abstract: Meta-analysis and empirical studies have reported that students felt discouraged and had lower motivation and higher test anxiety when taking Computerised Adaptive Tests (CATs) compared to fixed-item tests. Those negative experiences might reduce the validity of test score interpretation in CATs. Although the causes of negative experiences in CATs are still unclear based on the findings of meta-analysis studies, two possible explanations of the negative emotion in CAT were suggested as 1) high perceived item difficulty by Tonidandel et al. in 2002 and 2) quick increase of item difficulty (i.e., difficulty jumps) in the CAT testing process by Ortner and Caspers in 2024. Additionally, teachers monitoring the practical CAT situations reported that students perceived rapid increases in item difficulties between adjacent items at the beginning of the test. Therefore, based on the explanations above, assuming that 1) reduction of perceived difficulty and 2) difficulty jumps (i.e., changes of difficulties between adjacent items) can reduce negative experience, the current study manipulated CATs by adding 1) constraints of item difficulty never over ability estimates and 2) constraints of difficulty jump in a limited range in CAT. We explored the effect of the constraints on administered item difficulties, the extent of difficulty jumps, and measurement precision.

This study proposed and explored four constraints in CAT: Constraint of difficulty lower than ability (CDLA), Constraint item difficulty jumps (CDJ), Weighted Fisher information approach (WFI), and Fixed start item design (FS). The CDLA reflects the hypothesis of the effect of perceived difficulty, which constrains the difficulty of the selected item never over ability estimates. The CDJ and WFI reflect the hypothesis of the effect of difficulty jumps. The CDJ constrained the difference of difficulties between adjacent items in the testing process to lower than 0.5 logits. The WFI weighted the information function of the next processed item with a narrowed Gaussian distribution on the scale of item difficulty. The FS constrained the 12 start items in CAT as a fixed-item section with easy items to reflect the practical teachers' reports about difficulty jumps at the beginning of CATs. Compared to the traditional CAT as a baseline performance, simulation studies were conducted to evaluate the proposed constraints methods in terms of the administered item difficulties, changes of item difficulties in testing processes, and root mean squared errors (RMSE) of ability estimation.

Our results suggest that first, for middle-ability level test takers, the WFI can control the difficulty jumps with the least sacrifice of measurement precision (i.e., the least increase of RMSE). Second, for extremely high or low ability levels, CDJ would be recommended to successfully control difficulty jumps with the least sacrifice of measurement precision. Third, the CDLA can control the administered difficulty but run out of the item bank for low-ability test takers. A future study will discuss the connection between test-takers' negative experiences and the proposed constraints applied in practical CATs.

S5-4: Symposium- The development of iSKA

Chair: *Yongsang Lee*

Discussant: Hyun Sook Yi

Abstract: The King Sejong Institute Foundation has established 248 King Sejong Institutes in 85 countries as of 2023, delivering systematic Korean language education to approximately 120,000 students annually. In response to the rapidly increasing demand for Korean language learning worldwide, the Foundation initiated the Sejong Korean Language Assessment (SKA) in 2019 to evaluate and verify the proficiency of Korean learners comprehensively and accurately. The rising global interest in Korean language education underscores the importance of the SKA, which offers reliable and internationally recognized language proficiency assessment results.

Since 2022, the King Sejong Institute Foundation has been developing the iSKA (improved SKA), employing a multistage adaptive test (MST) to enhance the accuracy of the Korean language proficiency assessment and to increase efficiency by reducing the number of assessment items and testing time. The iSKA includes listening, reading, speaking, and writing sections, with listening and reading sections comprised of multiple-choice items, and speaking and writing sections using constructed-response items. The iSKA represents the first national-level test in South Korea to implement the MST format, leading innovations in testing and assessment. In 2022, prototype testing and simulations were conducted to explore the feasibility of the MST design for iSKA, and in 2023, further research was undertaken to refine the testing system, ensure the stability of its implementation, and establish a scoring system. The full implementation of iSKA is scheduled for 2024.

This coordinated session consists of four research, aiming to delve into the collective efforts and individual research studies that contribute to the development of the iSKA. Bringing together leading experts and researchers, this session will explore various facets of the iSKA development, from its theoretical underpinnings and design principles to the practical challenges and innovations in its implementation. The purpose of this session is to highlight the collaborative research dynamics that underpin the advancement of computerized adaptive testing in the context of Korean language assessment, demonstrating the potential impacts on educational technology and language proficiency testing globally. Through this discussion, we seek to share insights, foster dialogue, and encourage further research into adaptive testing methodologies that can enhance language learning and assessment internationally.

Presenters:

Yongsang Lee, Hwanggyu Lim, & Kyung (Chris) Han

Introduction to the Sejong Korean Language Multistage Adaptive Test (iSKA)

Abstract: The Sejong Korean Language Assessment (SKA) is designed to comprehensively measure the Korean language communication abilities of learners across four skills: listening, reading, writing, and speaking. Since 2022, the iSKA, an enhanced version of SKA based on a multistage adaptive test (MST), has been developed. This new design aims to improve the precision of ability estimates and tailor test difficulty to examinees' abilities, thus enhancing the testing experience.

The listening and reading sections of iSKA, composed of dichotomously-scored multiple-choice items, are developed based on the MST design. In contrast, the speaking and writing sections, composed of polytomously-scored constructed-response items, are developed as simple computer-based tests (CBT). These sections require a separate scoring process and are not suited for immediate score generation like adaptive tests. Consequently, this study focuses on the MST-based listening and reading sections.

The listening and reading sections are implemented using an algorithm based on the item response theory two-parameter logistic model (IRT 2PLM), ensuring stable estimation of item parameters and test-taker abilities in contexts with a limited number of participants. The MST design for these sections consists of a three-stage panel. At the first stage, a module of intermediate difficulty is presented. Initially, the second stage in 2022 included three levels of modules (easy, medium, and hard); however, subsequent research in 2023 led to the optimization of this stage by reducing it to two levels (medium-hard and medium-easy) to minimize the number of items used. The final stage consists of three modules, each differing in difficulty (easy, medium, and hard).

Each module in the MST panel contains 10 items, resulting in a total of 30 items for test-takers in each of the listening and reading sections. In the second and third stages, modules between adjacent difficulty levels share a few common items to minimize the total number of items required to construct the panel. Thus, the early 1-3-3 MST panel design, which required 70 items per panel, has been optimized to a 1-2-3 configuration, requiring only 54 items per panel through strategic design adjustments. Regarding the scoring system, abilities are estimated using the IRT 2PLM, and scores are subsequently converted to a standardized 200-point scale.

Overall, iSKA utilizes the MST design based on an item bank to provide test modules that match the learner's ability level during testing, minimizing measurement errors and maximizing the accuracy of Korean language proficiency assessments. Compared to traditional fixed-form SKA tests, iSKA significantly reduces the number of test items, alleviates the testing burden on learners, and

reduces the administrative load on the King Sejong Institute Foundation, thereby dramatically enhancing the convenience of Korean language proficiency testing.

Hwanggyu Lim, & Yongsang Lee

Advancing Language Assessment: Integrating Pretest Items and Automated Test Assembly in iSKA

Abstract: The Sejong Korean Language Assessment (iSKA) employs a multistage adaptive testing (MST) design and is currently implemented based on an item bank. However, being in its initial development phase, iSKA's item bank lacks sufficient items. Ensuring stable test administration and maintaining test security through minimized item exposure requires ongoing development of new items and expansion of the item bank.

Until now, iSKA has relied on costly and time-consuming separate pilot tests to estimate new item parameters. Integrating pretest items with operational items during main testing is now essential to gather student response data more efficiently. Furthermore, with the anticipated expansion of the item bank, transitioning from manual MST panel construction to automated test assembly (ATA) will enhance both the efficiency and accuracy of test assembly.

This study sets two primary objectives: Firstly, to design an MST panel that incorporates pretest items alongside operational items for more efficiently calibrating new items. This approach aligns with designs of several well-known adaptive tests, such as the Graduate Management Admission Test, which integrate pretest items to efficiently secure new items. iSKA aims to employ a similar strategy to robustly enhance its item bank.

Key research considerations for this design include determining the optimal number of pretest items and their placement within the test to minimize examinee burden while ensuring stable parameter estimation. For instance, including approximately 3 to 5 pretest items in Reading Assessment—about 10-20% of total operational items—balances manageable test length and reduces examinee fatigue. Final decisions regarding the number of pretest items will follow literature review and consultations with adaptive testing experts.

The placement of pretest items significantly impacts the accuracy of parameter estimation. Placing pretest items towards the end of the test can lead to student's guessing due to time constraints, compromising data reliability. Moreover, fixed positions of pretest items might disclose their locations, potentially resulting in disengaged responses. Therefore, this study will use experimental methods to identify the most stable seeding positions for pretest items under various conditions.

The second objective is to develop a roadmap for an efficient ATA system for constructing MST panels. With a growing item bank and the need to create multiple MST panels, it becomes unfeasible to manually satisfy numerous test constraints and psychometric properties required. Previous studies have demonstrated the effectiveness of using mixed integer linear programming (MILP) for assembling complex MST panels. iSKA intends to automate its test assembly using the MILP method for the ATA, exploring different

options for setting objective functions based on the psychometric goals of the test. For example, the target test characteristic curve may be used if optimizing test difficulty, while the target test information function is preferred for optimizing test information. Alternately, a combination of both functions can be set as the objective function. Simulation studies will be conducted to derive an optimized test assembly tailored to iSKA's design and objectives.

This research will enhance iSKA's capacity for managing adaptive testing environments effectively, providing a robust framework that ensures a reliable assessment of language proficiency.

Kyung (Chris) Han, Hwanggyu Lim, & Yongsang Lee

Framework for Score Scale Development for Multistage Adaptive Tests

Abstract: In the transition from traditional fixed test form designs to adaptive testing formats, such as Computerized Adaptive Testing (CAT) or Multistage Testing (MST), the effectiveness and appropriateness of existing scoring scales are often called into question. This shift necessitates a reevaluation of fundamental test properties, including scoring methods, test length, methods of theta estimation, reliability, and the conditional standard error of measurement. These properties of test may undergo significant changes, rendering previous scoring scales obsolete or not even usable. The case of the iSKA, an MST-based iteration of the Sejong Korean Assessment, exemplifies this challenge. Originally derived from a fixed test form, the iSKA prompted a critical examination of whether it was appropriate to carry over the established score scale. To address these concerns, a thorough discussion on the essential considerations for MST-based score scales was undertaken. This discourse led to the proposition of practical measures to assess the effectiveness and suitability of maintaining the existing score scale. The result of this process was the creation of a robust framework tailored for the development of new scoring scales that align optimally with MST-based assessments. The practicality and impact of this framework were demonstrated through its application in the real-world context of the iSKA.

This proposal aims to present the comprehensive framework and its application in iSKA as a case study, highlighting the nuanced considerations and methodological innovations that underpin the successful adaptation of scoring scales in MST-based testing. The insights gleaned from this experience are poised to inform and guide the development of adaptive testing programs, ensuring that they are equipped with scoring systems that accurately reflect the evolved testing dynamics and maintain the integrity of the measurement process.

Dongkwang Shin, Yongsang Lee, & Taehyun Kwon

Developing an Automated Scoring System for the iSKA Writing Test: Progress and Prospects

Abstracts: This study investigated the third year of work on an automated scoring program for the iSKA writing test. The primary objectives were to develop a linguistic feature analyzer to assist in human scoring and to create an automated scoring model. First, a tool was created to analyze 11 linguistic features. These features were divided into three groups: common features, task performance features, and language use features. However, an experiment demonstrated that statistical information of scoring features could not significantly contribute to traditional human scoring methods. To construct the automated scoring model, we combined writing responses from the November 2022 and 2023 iSKA exams. This yielded 1,235 responses for Question 1, 1,391 for Question 2, 1,096 for Question 3, and 904 for Question 4. The datasets were then divided into training, validation, and evaluation datasets [7:1.5:1.5]. The electra-kor-base model, a deep learning model trained on a large text corpus, was employed. While not requiring separate feature analysis, the model was further enhanced by incorporating selected information of the scoring features through an ensemble approach. The model initially demonstrated an accuracy range of 0.52 to 0.81, with an average of 0.6. After fine-tuning, the final model's performance improved, with an accuracy range of 0.67 to 0.82. This meets the reliability benchmarks for human raters. These findings suggest the potential for the development and application of future automated writing scoring systems.

S6-1: Paper Session- CAT applications 4

Chair: *Pradyumna Amatya*

Pradyumna Amatya

Transitioning from a linear to the adaptive foreign language testing: A case of DLIFLC's Computer Adaptive Defense Language Proficiency Test (DLPT-CAT)

Abstract: This presentation is about Defense Language Institute Foreign Language Center's (DLIFLC) Computer Adaptive Defense Language Proficiency Test (DLPT-CAT). In this presentation, the presenter outlines nine critical steps (Data extraction and preparation to Final cuts) in transitioning from a linear to the adaptive foreign language testing and discusses each step in detail. For instance, determining optimal CAT length via simulation and setting up pre-equating and post-equating studies. As an illustration, the presenter also provides simulated examples of DLPT-CAT administration at various theta levels (1, 2, and 3). The presenter ends his presentation by providing status of DLPT-CAT, lessons from current transitions, and plans.

Selma ŞENEL, Serkan ÇANKAYA, Hilal GENGEÇ, & Eren Can AYBEK

Computerized Adaptive Testing Design for Students with Special Needs: "Universal Test System"

Abstract: Collecting valid and reliable results from the tests applied to individuals with special needs is critical in terms of not putting on new obstacles to the existing disabilities of these students. For this purpose, accommodations and modifications are widely used in tests. However, literature presents major doubts about the validity, reliability and psychological effects of the tests applied with the accommodations.
"Universal design", which aims to enable all kinds of tools, products and designs to be used by individuals with different needs, has become a concept that is also used for tests. The purpose of this research is to design a computerized adaptive testing (CAT) system: "Universal test system", using "universal test design" principles.
Universal test system is aimed to provide an environment for the accommodations used in large-scale tests, where the individual with special needs can take the test independently without the need for an assistant, using up-to-date measurement approaches and methods. For this aim, a parametric computerized adaptive testing (CAT) software which can be easily used by individuals with special needs will be developed. Use of paper-pencil tests in special needs students require assistant assignment, additional time, special exam halls, additional hardware and special documentation. CAT may eliminate most of these requirements with fewer items, with high precision, as the student does not encounter items that are too low or far above his/her level with CAT, short duration, appropriate plug-in and computer assistive devices.
In this study, it is aimed to design such a CAT based Universal Design oriented Test System. Design-based research methodology will be used in this research. The views of subject experts and teachers will be taken in different periods during the software development. Pilot studies and views of the participants/experts/teachers will be carried out gradually and repeatedly. The software will be improved according to the results of the pilot studies and analysis of the qualitative data. After the software is developed, the main experimental application will be conducted with various special needs groups. The validity and reliability analysis of the test statistics will be made, and the data obtained from the interviews will be analyzed using content analysis.
The results of the research, a parametric and more accessible CAT system, will produce significant results for the literature and real life practices with its original, applicable and innovative features that combines measurement, special education and technology.

Mehmet Can Demir, Beyza Aksu-Dünya, & Stefanie A. Wind

Developing a Computerized Adaptive Test for Assessing Faculty Assessment Literacy

Abstract: Assessment literacy can be stated as an educator's technical knowledge and skills in assessment with substantial emphasis on psychometric principles and test design. Targeted professional development efforts in assessment literacy are essential to empower faculty to excel in their assessor roles. The purpose of this study was to generate an item bank for assessing faculty members' assessment literacy and, using this bank, to examine the applicability and feasibility of a Computerized Adaptive Test approach to monitoring assessment literacy among faculty members in higher education. In developing this assessment our main goal was to create a simple, quick, and precise screening tool related to assessment literacy in higher education. After carefully defining the construct of assessment literacy within the higher education context, the test blueprint and items were developed and underwent a series of expert reviews. Following a pilot administration to confirm feasibility, we carried out item parameter calibration with a representative sample (n=211) of faculty members from disciplines. In this calibration, we evaluated the items for psychometric quality, including fit, targeting, and unidimensionality under the Rasch methodology framework. Following the calibration, we conducted pilot CAT administration via FastTest (n=25) to test the efficiency of the item pool. We concluded that developing an adaptive test for measuring assessment literacy is possible even with a small item pool and a small calibration sample. The study concludes with practical implications for integrating adaptive assessment tools into faculty development programs, enhancing the efficiency and effectiveness of ongoing assessment literacy evaluations.

Haniza Yon

Development of An Adaptive Financial Knowledge Test For Loan Application

Abstract: Computerised adaptive testing (CAT) is a powerful technology-driven method in which a computer programme administers test questions according to a dynamic algorithm that continuously estimates the ability level of the test taker and chooses successive questions accordingly. An important consideration in development of any CAT is the size of the item bank that will be needed to estimate test-takers' ability with a satisfactory level of precision. This study examines the issue of item bank size and related parameters in the context of a financial knowledge test that is used to assess loan applicants' level of financial knowledge.

Monte Carlo simulation studies were carried out using the software package CATSim in order to explore combinations of item bank size and other parameters such as test length, item exposure, scoring algorithm and termination criterion that would be sufficient to determine loan applicants' scores with particular levels of precision. Once an appropriate precision threshold for the test had been selected, the results from the Monte Carlo studies were used to determine how many additional items needed to be developed and added to the item bank in order to meet that threshold. The required new items were then developed by experienced item writers. All items were pilot tested, and then calibrated using a dichotomous model. Post-hoc simulation studies incorporating real data from the pilot tests were also carried out to better predict how the adaptive financial knowledge test would perform with real loan applicants in the future.

S6-2: Paper Session- Multi-stage testing 1

Chair: *JP Kim*

JP Kim, Onur Demirkaya, Jing Ma, & Chris Han

Optimal Multistage Adaptive Testing Shaping Modules with Various Shaping Rules

Abstract: Multistage adaptive testing (MST) has gained gradual real-world adoption over a couple of decades for its unique advantages over linear testing. Among the various models, Han and Guo (2013) proposed MST by Shaping (MST-S) approach, which assembles a test module 'on-the-fly' after each stage by shaping the next stage according to the difference between the current and the target test information (TIF).
In MST-S, the test module shaping process is done iteratively, meaning that after the initial set of items for the next stage is randomly selected, the subsequent set of randomly drawn items is only chosen if they bring the existing TIF closer to the target TIF. Otherwise, the items are discarded. The process iterates until the predetermined number of items and iterations in the module is reached.
Their approach with a large enough number of iterations resulted in the smaller conditional standard errors of estimation (CSEE) than the traditional MST (MST-R) and could reach to the equivalent level of a-stratified CAT. The MST-S achieved measurement precision comparable to the MST-R after three shaping iterations in their studied conditions. With six iterations, its conditional mean absolute errors (CMAE) were close to the target and stable throughout the θ scale. MST-S effectively managed item exposure and utilized the entire item pool compared to other CAT and MST-R.
This study proposes and systematically compares three different types of loss functions for evaluating the area difference between the current TIF and a predefined TIF targets during the shaping process: (1) Mean Modified Parabola (MMP: effectively flagging even small difference: see Figure 1), (2) Mean Absolute Difference (MAD), and (3) Mean Squared Difference (MSD).
The simulation uses a 1-3-3 MST design as a baseline condition, with each module comprising 20 items, totaling 140 items. After each stage, an examinee's interim θ estimate is computed, and the next module is selected based on the maximum module information criterion. For exposure control, two additional panels with identical test characteristics are constructed, totaling 420 items in the MST-R setup. MST-R, Item CAT- maximum Fisher information (MFI), MST-S, MST-S with the absolute difference [MST-S(AD)] and target TIF stopping rule [MST-S(TS)] are compared.
The target TIFs for MST-S approaches are set for each stage and at three evaluation points on the θ scale: θ-1, θ, and θ+1. For the first stage, the TIF targets are 4, 5, and 4. For the second and third stages, they are 9, 15, and 9, and 12, 25, and 12, respectively. These targets ensure comparable precision as they are derived from MST-R module TIFs. The module-shaping process of three MST-S approaches examines three different conditions: 10, 100, and 1000 iterations.
Six thousand simulees randomly drawn from a U(-3, 3) are used in all methods. The initial θ value for selecting the first item (for CAT-MFI) is

randomly drawn from a U(-0.5, 0.5). Interim and final θ estimates are computed via MLE.
CSEE, CMAE, and conditional bias are computed to evaluate the measurement performance along with the item exposure and pool utilization.

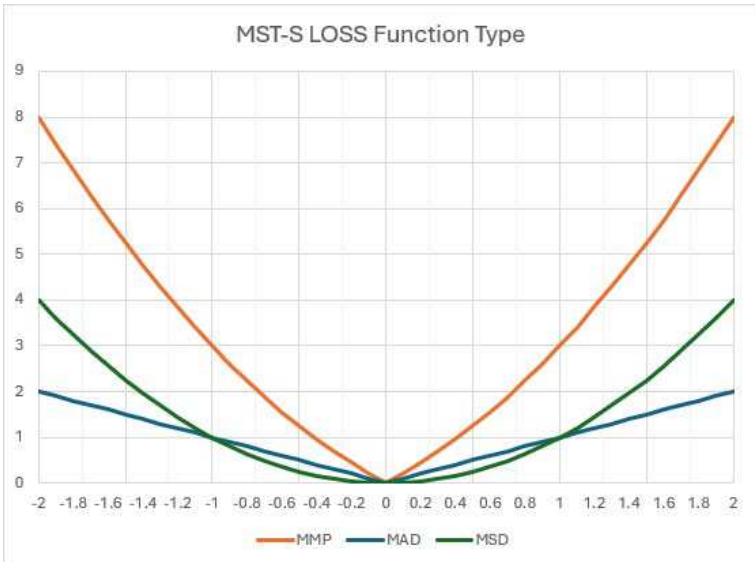


Figure 1. Comparison of the MST-S LOSS function across the three shaping rules proposed

Reference

Han, K. T. & Guo F. (2013). *An Approach to Assembling Optimal Multistage Testing Modules on the Fly*. GMAC® Research Reports, RR-13-01.

Insub Shin, Rae Yeong, Kim Sangyeon Jo, & Yun Joo Yoo

MathCoDi an online diagnostic assessment platform for teachers

Abstract: In offline classroom settings, teachers face several challenges in constructing and administering cognitive diagnostic tests solely. Before constructing a cognitive diagnostic test, it is crucial to configure a Q-matrix to elucidate the connections between test items and attributes. However, the process of identifying attributes for each item can be time-consuming for teachers, especially when teachers construct the test without others' support. Additionally, gathering a substantial amount of item response data is necessary to ideally estimate item parameters. Previous studies utilized around 1,000 item response data for estimation. But, in most classroom settings, there is an insufficient number of students available for such estimation purposes. Furthermore, analyzing the test results from the perspective of cognitive diagnosis theory and providing timely personalized feedback are also demanding challenges for teachers. MathCoDi, a mathematics online diagnostic assessment platform, offers an item pool and test assembly to assist teachers in addressing the challenges of constructing and administering cognitive diagnostic test. MathCoDi makes use of more than 300 cognitive attributes based on the Korean secondary school curriculum, facilitating the construction of assessments that align closely with the curriculum. Additionally, MathCoDi stores item-attribute relationships and item parameters for each item included in the item pool. This enables teachers to easily create cognitive diagnostic tests by selecting items from the item pool, eliminating the need to recruit a large number of examinees or manually configure a Q-matrix. Moreover, MathCoDi supports the design and execution of cognitive diagnostic multistage testing, delivering customized modules tailored to individual student's cognitive status. After the test is terminated, MathCoDi automatically scores student responses, interprets them based on cognitive diagnostic assessment, and visualizes them through a dashboard for both students and teachers. In MathCoDi's dashboard, the mastery levels of attributes measured through the test are visually represented in a graph structure depicting the hierarchical structure of attributes. Students can identify attributes of insufficient mastery on the dashboard and receive recommendations for learning paths to address those attributes. Teachers can review overall mastery information for their class across attributes on the dashboard, enabling them to adjust their instruction strategies accordingly.

Garrett Ziegler

Adapting SAT Reading and Writing for Multistage Adaptive Testing

Abstract: In 2020, the College Board began the most extensive redesign of the SAT in the nearly 100-year history of the assessment, moving from a model in which almost every student took the same paper-based test to a model in which each student takes a unique multistage adaptive test on the College Board's digital testing platform. By spring of 2024, the entire vertically scaled SAT Suite of Assessments (the SAT, the PSAT/NMSQT, the PSAT 10, and the PSAT 8/9), taken by more than 6 million students a year, had made this transition. Tests in the SAT Suite now consist of two sections—(1) Reading and Writing and (2) Math—that are each divided into two stages or modules, an initial routing module and a higher- or lower-difficulty second module to which students are routed on the basis of their performance on the initial module. For the Reading and Writing section of the test, the redesign was particularly extensive, with a nearly 50% reduction in the number of scored items (while maintaining the preexisting score scale) and a change from an entirely set-based assessment to an entirely discrete-item assessment. In this presentation, I will provide an overview of the changes to the test with a particular focus on the Reading and Writing section, discussing the overall design of the test (including information about scoring), how we identified which specific literacy-related skills can and should be assessed in this new format, and how the shift to multistage adaptive testing has altered aspects of test development and pool planning for the Reading and Writing section. The presentation will also touch on potential uses of generative AI to enhance the item-development process for SAT Reading and Writing, particularly with regard to adjusting task difficulty to meet the needs of adaptive testing.

Kyoungwon Bishop, Hacer Karamese, & Sooyong Lee, & Syed Hadi

Comparing Vertical Scaling Methods in Multistage Adaptive Testing using Simulation and Operational Data on WIDA's Language Assessments

Abstract:

Objective: This study investigates the effectiveness of concurrent and separate calibration methods for establishing vertical scaling in WIDA's multistage adaptive system (MST) language assessment in listening and reading domains. The research compares these calibration methods using both simulated and operational data to identify the strengths and limitations of each approach within WIDA's unique growth patterns and MST framework.

Background: Vertical scaling is crucial in large-scale language assessments as it enables the tracking of students' academic progress over time by placing test items across various grades on a common scale. WIDA's ACCESS language test is organized into grade clusters (grades 1, 2-3, 4-5, 6-8, 9-12). WIDA's vertical scaling employs a common item linking design between clusters, where 20% of the lower cluster's operational items are administered as vertical linking items in field test slots in the upper cluster test.
Vertical scaling in WIDA's MST design presents unique challenges, including the growth pattern of academic English across grade clusters and the tier assignment of vertical anchor items in the MST system. Each vertical scaling method has its advantages and challenges, necessitating a thorough comparison to determine the best approach for MST environments, particularly for language assessments with unique growth patterns and item assignment designs.

Methods: The vertical scaling simulation study was conducted to address issues related to growth patterns and MST systems using the Rasch model. Six simulation conditions were created, varying sample size, representation of tiers, and theta values to examine recovery rates of parameters and theta. Concurrent calibration estimates parameters for all items across grade clusters simultaneously. Separate calibration using the mean/mean method involves calibrating each cluster test individually and linking the clusters using a chained approach, with cluster 4-5 as the base. The calibration results of simulated and operational data were compared by examining cluster-to-cluster growth, variability, and separation of distributions, considering WIDA's unique growth patterns, where English learners' scores increase until cluster 4-5 and then decline in clusters 6-8 and 9-12.

Results: The analysis revealed both similarities and differences in ability and item difficulty estimates across clusters for the calibration methods between simulation and operational data. Simulation data exhibited greater variation than operational data, and initial parameters shifted in the operational data. Overall patterns of growth and difficulties were similar in both methods. However, separate calibration results showed a more significant decrease in growth in in growth in Listening cluster 6-8 and a slightly more increase in Reading cluster 6-8 compared to concurrent calibration. In the listening domain, differences were noted between simulated and operational analyses in separate calibration regarding item parameter recovery and ability

estimation, showing a greater distance from the simulated results with more items flagged.

Significance: The findings will guide the selection of the most appropriate calibration method for WIDA's operational vertical scaling, ensuring accurate monitoring of students' academic growth over time. The results of this research will have implications for other large-scale assessments employing MST systems and seeking to establish vertical scales.

S6-3: Paper Session- Software & Systems related to adaptive testing

Chair: *Chae Eun Kim*

Chae Eun Kim, Hyeon Park, Seong Eun Kwon, Jeongwook Choi, & Dong Gi Seo(p)

Development of LIVECAT Platform implementing Computerized Adaptive Testing (CAT, MST) in South Korea

Abstract:

Introduction: The LIVECAT platform is a testing system that implements Item Response Theory (IRT) models to provide computerized adaptive testing. LIVECAT calculates the results of each test using two scoring methods: Maximum Likelihood Estimation (MLE) and Expected a Posteriori (EAP). These features demonstrate superior performance compared to traditional testing platforms.

The process of implementing a test on LIVECAT: The LIVECAT uses several item response model as (1) Rasch model, (2) one-parameter logistic model, (3) two-parameter logistic model, and (4) three-parameter logistic model. Before a specific examination is formed, the administrator registers item category, item type, IRT model, item parameter estimates, and the number of answer options in the item bank. Items are not selected directly from the item bank during the examination. The administrator makes an item pool for each specific test.
Basic information should be added to construct a test, such as the title of the test, the start and stop time, the maximum and minimum item number and the examinee's information. Then, CAT components, such as scoring method, item selection method, termination criterion must be set.
The LIVECAT provides MLE as the default scoring method. The LIVECAT platform uses the Maximum Fisher Information (MFI) method as an item selection method. The method selects the next item that maximized information given the examinee's ability. LIVECAT provides three options for test termination criteria: maximum number of items, standard error of estimates and test period.
Test administrators set the test URL in the construct section for examinees to access. Examinees confirm their ID and password to take the test. Test results including true score, transformed t-score, and pass/fail status, are available immediately after completion. LIVECAT allows the downloading of examinee resposnes, ability estimates, and item codes in spreadsheet format.

CAT advantages: The LIVECAT platform can enhance test security by providing different sets of items to individual candidates based on their proficiency levels, thereby reducing item exposure and decreasing misconduct among candidates. Additionally, it allows for efficient and accurate measurement of latent traits using a small number of items, leading to cost savings through reduced creation of new items.

Comparison to traditional methods: The LIVECAT is accessible with just a few days of training for individuals familiar with computer usage, regardless of their expertise in psychometric measurement. Furthermore, the platform is

built on an IRT system, which allows for easy integration of additional features like MST and enables swift adaptation to industry requirements.

Conclusion and expected results: The LIVECAT platform, developed first in Korea and presented in the Korean language, facilitates ease of use for psychometric professionals but also for individuals without proficiency in English. In the near future, the LIVECAT will be updated to include features such as polytomous item response models, weighted likelihood estimation method, MAP, and content balancing method, further enhancing its capabilities.

Mayank Kumar

Probabilistic model for CAT: Administering tests without exposing answer keys

Abstract: At present, Computer Adaptive Testing (CAT) currently uses pre-calibrated items based on a large sample of responses, which require prior knowledge of the items difficulty level and discrimination parameter. There are two problems with this approach. Firstly, for new items whose discrimination parameter is unknown, pilot testing has to be done. Secondly, many institutions use third-party service providers to conduct exams. In order to maintain confidentiality, these institutions may not want to share the answer keys with external parties.

These problems can be solved by applying our probabilistic model-based application. In this model, there are two phases namely Test creation phase and Test administration phase. In test creation phase, the author will compose the test items after which along with a reviewer, the author will specify the degree of difficulty as well as the answer key. After the author has finished writing the test, our application will generate a "Set" for each of the item. A "Set" is a collection of one of the distractor and the answer key. After a set is created, our code encrypts the test and removes the answer keys that the author had previously provided. During the test administration phase, the application at the candidate's end decrypts the test.

In traditional CAT methods, only the current item answered by the candidate is evaluated in order to determine the difficulty level of next question. In our model, two to three consecutive items are analyzed. If a candidate selects any of the option that is present in the set, consecutively for two to three items, the candidate will advance to item with higher difficulty, else the candidate will get item of a same or a little lower difficulty.

The reason it is called a probabilistic model is because even though the candidate has an advantage of selecting the distractor or non-key from the set, in order to move to items of higher difficulty, the probability of that happening is 16% for every two to three items.

For example, for each question a set contains 1 distractor (out of 4 distractors) and 1 answer key.

Probability of selecting one of the options from that set = 25 or 40%

Now, in order to move to next difficulty, candidate needs to select one of the options from the set of that question for 2 consecutive questions,

Therefore, Probability of selecting one of the options from the set for 2 questions consecutively = 25×25 = 0.16 or 16%

S6-4: Paper Session- Machine learning

Chair: *Nathan Thompson*

Nathan Thompson, Sean Gasperson, Heidi Banerjee, & Fernando Austria Corrales

Developing a Custom Machine Learning Algorithm to Score Essays for a High-Stakes Educational Equivalency Assessment

Abstract: The use of human marking of open-ended items, while effective, can present challenges to the scoring of tests such as delays in score reporting, lack of standardization across raters, and the resources required to scale up test taker volume. Implementing machine learning in the scoring process presents an opportunity address these challenges. However, it is critical to undertake a rigorous process that adheres to best practice when developing the algorithm to ensure that test takers are evaluated fairly and accurately, and that no loss of quality occurs during the implementation. This study will provide an example of the development of such an algorithm using real-world test taker data, including the steps taken to iterate and improve the models using new responses.

We will utilize a training/test set approach, as is the standard for calibration of machine learning models. This approach splits the sample, trains a model on one portion, and then cross-validates on the other portion to evaluate generalizability. We will evaluate several percentages of splits (50/50, 70/30, 90/10). We also will investigate many types of features, including immediate text-based features such as the document-term matrix (bag-of-words model), and non-text features such as average words per sentence or Flesch reading ease index. Multiple models will be fit, including neural networks, random forest, and logistic regression. We will evaluate the models with human-AI score correlation and agreement.

Participants are test takers for PSI's HiSET exam. We will use their responses on multiple essay prompts, the rubric for grading their responses, and any scores given by the human evaluators. We will be using operational data from historical responses to various essay prompts. Currently, essay items are recorded, then scored electronically by trained human markers with specific rubrics for each essay item.

We will use a portion of responses on a specific prompt, the prompt, rubric, and human ratings to create multiple machine learning models. Then, we will use more responses on the same prompt to produce scores using the various models, and compare these scores with the human ratings. We will repeat this process using various divisions of the dataset, e.g., 50%, 70%, 90%, to evaluate which method produces the best results (i.e., comparability with assigned scores). Finally, we will apply the model to different essay items to determine its applicability across various prompts.

Lihua Yao, Aaron James Kaat, Catherine Han, & Richard Gershon

Machine Learning for Gaze Location Prediction for Gaze Measure in NIH Baby Toolbox

Abstract: The NIH Infant and Toddler Toolbox (commonly referred to as Baby Toolbox or NBT) is a specialized measurement tool designed for children aged 16 days to 42 months. This comprehensive toolkit, tailored for administration via iPad, encompasses a range of assessments aimed at evaluating critical developmental domains. One notable feature of the NBT is its incorporation of built-in eye-tracking technology, which facilitates the presentation of stimuli such as short animations or images. This technology enables the precise capture of a participant's gaze towards specific objects, allowing for the collection of valuable data on objective looking behaviors, attention accuracy, reaction time, and habituation speed.Central to the scoring process is the variable known as "gazeLocationName." This variable holds significant importance, as it plays a key role in determining scores. However, deriving the values of "gazeLocationName" is a complex task, with numerous instances where values cannot be accurately derived. Given the critical nature of "gazeLocationName," our study employed machine learning techniques to predict those with missing values. By leveraging features derived from the X and Y coordinates of the gaze, we aimed to predict "gazeLocationName" with precision. Our analysis revealed promising results, with an accuracy rate exceeding 70% for both test data and prospective future data. Additionally, the Area Under the Curve (AUC) metric surpassed 0.76, indicating a high level of predictive performance.

Young Jin Kim & Jin Eun Yoo

Utilizing collateral information with machine learning in the cold-start problem in computerized adaptive testing: A Monte Carlo simulation study

Abstract: In computerized adaptive testing (CAT), the cold-start problem often arises due to the lack of information about an examinee's latent trait at the beginning of the test, resulting in delays in providing personalized test items. While previous research has attempted to use collateral data to estimate empirical prior information, the scope of the collateral information was limited and did not consider prediction or generalization from the perspective of machine learning (ML). To address the cold-start problem, we propose using penalized regression, an ML technique, which is particularly suitable for CAT of readily available collateral information. Penalized regression also offers the advantage of producing interpretable prediction models.
We conducted two Monte Carlo simulation studies with 1,000 examinees and 200 collateral information variable. Simulation 1 compared the prediction performance of four ML techniques (LASSO, elastic net, Mnet, random forest) across three signal-to-noise ratio (SNR) levels (0.25, 0.5, 1). Based on the results, Simulation 2 assessed measurement precision and test length, evaluating 36 condition combinations: three SNR levels (0.25, 0.5, and 1), two initial item levels (difficulty parameter close to zero or the predicted level of each latent trait), three latent trait estimation methods (maximum likelihood, maximum a posteriori (MAP) with a standard normal distribution, and MAP with empirical prior), and two termination criteria (fixed length, variable length).
In Simulation 1, penalized regression techniques (LASSO, elastic net, Mnet) tended to outperform random forest in terms of prediction, with Mnet being the most parsimonious. In Simulation 2, the SNR turned out to be crucial for leveraging empirical prior information. Specifically, when the SNR was 1, using Mnet estimates as empirical prior information was effective. Furthermore, the fixed-length CAT with Mnet for latent trait estimation showed significantly higher measurement precision than the other conditions, and its variable-length counterpart reduced test length by between a minimum of 47.7% (7.7 items) and a maximum of 54.2% (10.3 items) compared to the other conditions. When the SNR was 0.25, on the other hand, the estimation with empirical prior information resulted in significantly lower precision than MAP with standard normal distribution.
In conclusion, penalized regression outperformed random forest in CAT with collateral information, and the level of SNR appears to be a crucial factor in utilizing empirical prior information to address the cold-start problem. Further research is warranted on adding strong predictors of a trait such as item response tiem and previous test scores to the model. Also, this study was an item-level CAT, whereas multistage testing have been actively studied in recent years. Therefore, it is necessary to examine the effect of applying the method proposed in this study to multistage testing. Finally, although this study attempted to emulate real data, there may be limitations. Therefore, live CATs with real examinee's are required. (455 words) Strand: CAT applications

S7-2: Paper Session - Item Selection Methods 2

Chair: *Rodrigo S. Kreitchmann*

Rodrigo S. Kreitchmann, Miguel A. Sorrel, Diego F. Graña, & Francisco J. Abad

Breaking down estimator variance: ipsative versus normative information in IRT models.

Abstract: Ipsativity is an important concern in psychological assessment, particularly with forced-choice response formats. It refers to the lack of comparability of test scores between persons, providing information only about the predominance of traits within a person and little or no information about the person's absolute standing in each trait. It originates from the multicollinearity between trait scores. For instance, in forced-choice formats, endorsing statements associated with a given trait implies not endorsing statements from the remaining dimensions, resulting in a negative interdependence between scores.
Ipsativity can largely affect score validity, as the sum of covariances with external variables and the sum of the trait variance-covariance matrix are biased towards zero. In practice, the composite/average scores across all traits tend to be constant, so the absolute position of examinees is unidentified. In the context of CAT, it is important not only to develop item selection methods that can reduce score ipsativity but also that can be able to quantify the ipsativity of the final scores.
This study proposes a way to derive the trait estimator variance into a) the covariance between trait estimators and true person vector average, and 2) the covariance between trait estimators and differential scores respective to the person average. In other words, it aims to quantify the variance in the scores that is attributable to the true composite variance, cov(,), and the one associated with the differential scores respective to the within-person average, cov(,). A high amount of true average variance indicates precision in estimating the absolute standing on the trait continuous, whereas the opposite can serve as an indicator of remnant ipsativity.
A simulation study was conducted to illustrate how the two proposed variance decomposition methods are estimators of the actual the covariances between 1) estimated scores and true composites, and 2) estimated scores and true differential values respective to the vector mean. Additionally, a follow-up simulation compared the use of those in item selection rules. A five-dimensional item pool of 240 items was simulated, and forced-choice pairs were generated from any non-unidimensional combination of the items in the pool (i.e., 23,040 possible pairs). CAT length was fixed to 30, and three item selection rules were assessed: a) maximizing cov(,), b) maximizing cov(,), and c) the A-rule. True trait correlations were either generated as zero or those (positive) found in the NEO Personality Inventory-Revised validation study. The trait score recovery in each condition was compared in terms of squared correlations between true and estimated trait scores and the recovery of the correlation with a simulated criterion.
As results, maximizing cov(,),) outperformed the A-rule, especially regarding the recovery of the correlation with a simulated criterion. On the contrary,

maximizing cov(,) gave place to ipsative scores, with very small amount of true information regarding the absolute standing of the scores, despite the very precise estimate of the relative predominance of the traits. These results evidence not only the utility of the two new indices to quantify ipsativity, but to improve item selection with forced-choice data.

Jinha Kim, Dong Gi Seo, & Jeongwook Choi

Comparison of real data and simulated data analysis based on the standard error of measurement for stopping algorithm in a computerized adaptive testing

Abstract:

Background: In computerized adaptive testing (CAT), the objective is to ensure consistent measurement precision across all examinees, regardless of their latent trait levels. Various stopping rules, including the standard error of measurement (SEM), have been proposed to achieve this in unidimensional CAT. The SEM, particularly, is crucial as it determines when to cease item administration based on the desired precision level.

Objective: The objective of this study is to identify the optimal stopping rule in computerized adaptive testing (CAT). Specifically, the study seeks to determine the ideal standard error of measurement (SEM) that should be used as a termination criterion to ensure an efficient trade-off between the accuracy of the estimated abilities and the efficacy.

Methods: This research is divided into two distinct studies: an empirical study and a post-hoc simulation study.

Empirical Study using LIVECAT Platform: The empirical study was conducted with data from a 2020 medical examination at Hallym University College of Medicine. A total of 1,012 item parameters, estimated using the Rasch model from the assessment data, were input into the LIVECAT platform to construct the computerized adaptive test (CAT). The CAT examination was administered to 83 medical students at Hallym University. The score calculation was performed using the Maximum Likelihood (ML) estimation method. Two SEM thresholds were applied as stopping criteria: SEM=0.25 and SEM=0.3.

Post-hoc Simulation Study: The post-hoc simulation study utilized the same item parameters estimated from the 2020 clinical assessment data. For this simulation, 1,000 virtual examinees were generated, with ability levels drawn from a standard normal distribution, N~(0,1). The CAT simulations were executed using both Maximum Likelihood (ML) and Expected a posteriori (EAP) estimation methods to calculate scores. The accuracy of the examinees' standard scores was evaluated at multiple SEM levels as stopping criteria: 0.1, 0.15, 0.20, 0.25, and 0.3. The purpose of the simulation was to assess the influence of different SEM thresholds on the balance between measurement precision and the number of items administered, thereby determining the ideal SEM for use as a termination criterion in CAT settings.

Results: The analysis of the results from both the empirical study and the simulation indicates a clear trend related to the standard error of measurement (SEM) thresholds in computerized adaptive testing (CAT). As anticipated, a decrease in the SEM threshold correlated with diminished measurement accuracy and a reduced number of items administered. This inverse relationship highlights the trade-off between the precision of the estimated abilities and the length of the test. Notably, the findings from both the real data set and the simulated scenarios were congruent. A SEM of 0.25

emerged as the optimal point, which maintained a high level of measurement accuracy while also minimizing the test length, thus supporting its viability as a practical termination criterion.

Um i Lela

Optimizing Item Selection and Termination Criteria: Integrating Human Expertise and Computational Approaches in Computerized Adaptive Testing

Abstract: Computerized adaptive testing (CAT) has revolutionized the way psychological and educational assessments are conducted, offering efficiency and personalization. However, the design of effective CAT systems requires a careful balance between human expertise and computational power. This presentation will explore the synergistic integration of human judgment and algorithmic approaches in the critical areas of item selection and termination criteria.

The first part of the presentation will focus on item selection methods, highlighting the strengths and limitations of purely computational approaches, such as maximum information, minimum expected posterior variance, and sequential probability ratio testing. While these algorithms can optimize item selection based on statistical criteria, they may overlook important content-related, cultural, or contextual factors that can greatly impact the fairness and validity of the assessment.

To address this, the presentation will introduce a hybrid approach that combines human expert input with computational item selection. Subject matter experts, test developers, and experienced assessors will be engaged to provide their domain-specific knowledge and insights, which can then be incorporated into the item selection process. This integration of human expertise and algorithmic optimization will be discussed, along with case studies demonstrating its benefits in terms of improving content coverage, reducing bias, and enhancing the overall test-taker experience.

The second part of the presentation will delve into termination criteria in CAT. Traditional approaches, such as fixed-length tests, variable-length tests with a stopping rule based on measurement precision, and tests that continue until a predetermined level of statistical confidence is achieved, will be reviewed. However, the presentation will also highlight the potential limitations of these purely computational termination criteria, particularly in terms of their impact on test-taker engagement, assessment purpose, and holistic evaluation of competence.

To strike a balance, the presentation will propose a collaborative framework that leverages both human judgment and computational analysis to determine appropriate termination criteria. Factors such as content coverage, construct representation, and overall assessment goals will be incorporated, along with statistical considerations, to ensure a comprehensive and contextually relevant approach to test termination.

Throughout the presentation, empirical research and practical case studies will be used to illustrate the implementation and performance of the proposed hybrid approach, highlighting the benefits of integrating human expertise and computational power in the design of effective and equitable CAT systems.

S7-3: Paper Session - CAT in Reality

Chair: *Arvind Singh*

Arvind Singh, Prashant Kapoor & Shivam Bohra

Adaptive Diagnostic Assessments: A novel method to diagnose learning gaps using a personalized adaptive approach

Abstract: Diagnostic assessments play a crucial role in evaluating proficiency levels on various competencies and identifying areas of improvement in learning. Traditional diagnostic assessments often rely on static, one-size-fits-all assessments that fail to account for the unique learning needs and abilities of individuals. To address this limitation, we propose a novel approach for creating adaptive diagnostic assessments that adapt to the individual's specific abilities and learning trajectory.

Our approach leverages item response theory and learning analytics techniques to develop personalized assessments that dynamically adjust the difficulty and content of questions based on the individual's responses. The backbone of the assessments is high quality test items which have been created with rigor and are able to bring out gaps in student understanding. Key elements include the utilization of item response theory models to estimate the learner's ability level and item difficulty parameters, and the incorporation of adaptive algorithms to select and present questions in real time tailored to the individual's proficiency. We propose an adaptive logic based on balancing item difficulty and student latent ability before each item is surfaced. We also propose an innovative item selection strategy that balances the need for accurate estimation of ability levels while minimizing assessment length.

Xiaowen Liu

The Impact of Missing Data on Parameter Estimation: Two Examples in Computerized Adaptive Testing

Abstract: Computerized adaptive testing (CAT) can provide more efficient and accurate estimation of proficiency compared to linear tests (Magis & Barrada, 2017). Because of this efficiency, many educational applications, including high stakes admission testing, large scale international assessments, and education training tools use some kind of adaptive.
Data generated by a CAT has a substantial amount of missingness relative to traditional testing designs because each examinee usually sees a very small subset of the item pool (Jewsbury & van Rijn, 2020, Magis et al., 2017). Inference regarding item and person parameters in the presence of missing data in CAT has been discussed with reference to Rubin (1976) general theory of inference with missing data. Mislevy & Wu (1996) pointed out that adaptive testing designs lead to ignorable missing data when the missingness depends on the prior observed responses. The ignorability of adaptive testing designs, such as multistage testing designs (MST) and standard CAT, has been studied to evaluate the quality of likelihood-based estimation (Eggen & Verhelst, 2011; Mislevy & Wu, 1996; Glas, 2009, Wang et al., 2020).
Jewsbury and van Rijn (2020) conducted an in-depth study to show that certain operational practices might lead to biased parameter and ability estimation. Specifically, they noted that some large-scale assessments administered a multidimensional multistage tests (MST), but then analyzed the data separately by dimension. If the data were analyzed as a multidimensional IRT model (mirt), the missing data would be "missing at random" - or MAR in the (Rubin, 1976) terminology - because the missingness mechanism depends only on the observed responses from the routing stage. However, if each dimension is analyzed separately as a unidimensional model, then the omitted responses to items from other dimensions render the missingness MNAR (missing not at random). They showed substantial bias in the estimated item discriminations and intercepts.
The objective of the current study is to investigate the impact of missing data in computerized adaptive testing by examining person ability parameter and item parameter estimates in various test conditions that satisfy and do not satisfy MAR assumption. Specifically, we used three examples yielding both acceptable and unacceptable results when using post-CAT data for parameter re-estimation. First, we investigate recalibration using post-CAT testing data. We simulate test administrations from a large item pool where the respondents either saw all the items, a random set of 30 items, and a 30 item CAT. Parameter estimation in the CAT can be surprisingly good despite the challenges of the sparse data. Second, we showed that there can be unusual patterns in the estimates where discrimination parameters swing negative for the high difficulty items. We explored this result which was reproducible in certain CAT designs. Finally we showed a multidimensional CAT where the MAR assumption was violated and caused the negative slopes to happen regularly.

Practitioners who are interested in utilizing post-CAT data for further analysis need to unravel whether the process of re-estimation post-CAT data

violated the MAR assumption, and choose correct models for the re-estimation.

David Budzyński, Matthew Turner, Ben Smith, Andrew Boyle, & Sefa Sahin

When Adaptivity Meets Reality: Addressing User Concerns and Pedagogical Challenges

Abstract: AlphaPlus leads the consortium for developing Online Personalised Assessments (OPAs) in Wales [1], powered by Cito's algorithm [2]. They replace Welsh National Tests (WNTs), offering on-demand, formative assessments with next-day feedback. OPAs use a common item response theory scale across eight school years, ensuring content matches learners' abilities and avoiding floor/ceiling effects. The first assessments launched in 2018, with the first National Report on attainment released recently [3].

As consortium lead, we prioritize continuous improvement. Feedback drove enhancements to procedural numeracy assessments in 2022-23, with a redesigned version in 2023-24, well received. This prompted similar improvements to other assessments. However, Welsh and English reading assessments, started in 2019, pose a unique challenge. They involve multiple questions per text, requiring the adaptive algorithm to select suitable question sets rather than adapting item-by-item.

One of the most interesting pieces of feedback challenged the "purist adaptive" model of these assessments. This has sparked discussions about the nature of reading versus numeracy assessments and the explicit teaching of a concept required before assessment [4]. There is also concern about how many constraints can be applied before an adaptive assessment becomes akin to a traditional linear assessment, reducing its adaptive value [5].

Our responsibility is to balance user feedback with the need for a psychometrically robust instrument. Currently, we are piloting adjusted designs for the reading assessment with these changes:

1. Reducing the curriculum content targeted at years above the learner's level to address concerns about complex texts being inaccessible to younger learners.
2. Implementing a more defined "ramping" of difficulty, allowing learners to ease into the assessment without a sudden increase in difficulty.
3. Allowing "branching paths" to provide adequate challenge for the most able learners while still offering opportunities for all learners to perform their best.

These pilots will conclude in May 2024, enabling us to share findings and confirm the final design for launch in September for the new academic year. In this presentation, we reflect on the tension between "pure" adaptivity and practitioner perspectives on assessment operation, detailing the process of reaching a compromise. We aim to share our insights for other systems considering or adopting adaptive assessments, and generate a discussion within the adaptive testing community about the best ways to achieve a balance between psychometrics and pedagogy/user expectations.

References

Welsh Government. Information for parents and carers of children in Years 2 to 9 in maintained schools in Wales. Accessed: 2023-06-04. 2023. url: https://hwb.gov.wales/api/storage/d97df63f-ea36-48cd-ab6d-934df

89b438f/online-personalised-assessments-in-readingandnumeracy-years-2-to-9-in-maintained-schools.pdf.

Angela J. Verschoor. Cito CAT engine. Software for Computerized Adaptive Testing. Arnhem: Cito, 2024.

Welsh Government. Patterns in reading and numeracy attainment: from 2018/19 to 2022/23. Accessed: 2023-06-04. 2023. url: https://www.gov.wales/patterns-reading-and-numeracy-attainment 2018-19-2022-23.

Walter Kintsch. Comprehension: A Paradigm for Cognition. New York: Cambridge University Press, 1998.

Karen A. Becker and Betty A. Bergstrom. “Test Administration Models”. In: Practical Assessment, Research & Evaluation 18.14 (2013). Accessed: 2023-06-04. url: https://pareonline.net/getvn.asp?v=18&n=14.

S8-1: Paper Session - Multi-stage Testing 2

Chair: *Hanan AlGhamdi*

Hanan AlGhamdi, Gorgeous Sideridis, & Omar Zamil

Contrasting Multistage and Computer-Based Testing: Score Accuracy and Aberrant Responding.

Abstract: The aim of this study was to evaluate the effectiveness of a multistage testing (MST) design with three pathways compared to a traditional computer-based testing (CBT) method, which includes items for all ability levels. Specifically, the research sought to compare and contrast the efficacy of these two testing approaches in terms of their ability to provide accurate person ability estimates and detect aberrant responding patterns.

Participants in the study consisted of 627 individuals who were subjected to both types of assessments. The CBT instrument comprised a comprehensive set of items designed to measure abilities across a broad spectrum, ensuring that all ability levels were adequately represented. In contrast, the MST design utilized a routing mechanism that directed individuals to one of three paths—low, middle, or high ability—based on their initial responses, thereby tailoring the content to better match each participant's ability level.

The comparison between MST and CBT focused on several key metrics. One primary area of interest was the accuracy of the ability estimates produced by each method. Another critical aspect was the evaluation of aberrant responding, which refers to patterns of responses that deviate from expected behavior, potentially indicating guessing, carelessness, or other irregularities. Person-fit statistics were used to identify these aberrant response patterns.

The results of the study indicated notable differences between the MST and CBT assessments. Specifically, MST assessments showed a marked deviation from CBT assessments, particularly for individuals at the lower and higher ends of the ability spectrum. Overall, test score accuracy was higher in the MST design compared to the CBT approach, suggesting that the tailored routing mechanism in MST was more effective at accurately assessing individuals' abilities.

However, the study also found that the error of measurement was greater for high-ability individuals during MST compared to CBT. This finding suggests that while MST generally enhances accuracy, it may also introduce some degree of measurement error for certain subgroups.

Further analysis of response patterns revealed significant differences in the incidence of Guttman-related errors, which are indicative of inconsistencies in response behavior. The CBT method exhibited a higher frequency of these errors compared to the MST approach, as evidenced by the person-fit aberrant response indicators.

In conclusion, the study found that MST offers significant benefits over traditional CBT. The tailored approach of MST not only improves overall test score accuracy but also reduces the occurrence of aberrant responding. These findings highlight the potential advantages of MST in educational and psychological assessments, particularly for more accurately measuring abilities across diverse populations.

Jinmin Chung, Sungyeun Kim, Jaehwa Choi, & Dayeon Lee

Revolutionizing Multi-Stage Testing: A Comparative Study of Template and Rule-based, Language Model Only, and Ontology-Model Centered Generative Approaches

Abstract: Multi-stage testing (MST; Drasgow, et al., 2006) represents an advanced form of computerized adaptive testing (CAT; Wainer & Kiely, 1987; Lewis & Sheehan, 1990; Sheehan & Lewis, 1992; Wainer & Lewis, 1990), offering structured, modular test stages that adapt to the test-taker's performance. While CAT dynamically selects items for each examinee in real-time, MST pre-assembles sets of testlets (modules) that adapt based on the test-taker's ability at predefined stages. This structured adaptivity provides greater control over test content and allows for more complex test designs.

The evolving needs of MST have necessitated the integration of generative approaches to enhance the adaptability and precision of test construction. Traditional Template and Rule-based (TR) generative MST (GMST; Choi, 2024) relies on predefined rules and heuristics to generate and adapt test modules. In contrast, Language Model Only (LOG; Choi, 2024) GMST leverages advanced large language models (LLMs; Brown et al., 2020; Choi, 2024; Kojima et al., 2022; Vaswani et al., 2017) to dynamically generate test content and structure based on vast amounts of training data, offering a more flexible and scalable solution.

Additionally, ontology-model centered Generation (OMG; Choi, 2024) MST emerges as a hybrid approach, combining the template-based structure of rule-based systems with the generative capabilities of LLMs. In OMG MST, templates and ontologies guide the structure and content of test items, ensuring alignment with educational standards and specific learning objectives. LLMs are employed within this framework to dynamically generate and adapt test items, providing contextual richness and diversity while maintaining coherence with the predefined templates.

This study compares TR, LOG, and OMG GMST approaches. TR GMST provides clear, interpretable rules and ensures compliance with specific test design principles, but may lack flexibility and require extensive manual input to cover all potential scenarios. LOG GMST can generate diverse and contextually rich test items, adaptively refining the testing process in real-time, though it may face challenges in interpretability and potential biases from the training data. OMG GMST bridges these approaches, offering a balance between structure and flexibility, leveraging the strengths of both TR and LOG methodologies.

The comparative analysis reveals that while TR GMST offers reliability and transparency, and LOG GMST significantly enhances adaptability and efficiency, OMG GMST provides a balanced, scalable solution that combines the precision of rule-based systems with the dynamic capabilities of modern AI-driven methods. The findings underscore the potential of OMG GMST to revolutionize MST, delivering both structured precision and adaptive flexibility in test generation and execution. All three approaches were implemented using the CAFA (Collective AI on the Foundation AI; Choi, J., Kim, S., & Yoon, K., 2004) platform.

Reference

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems, 33*, 1877-1901.

Choi, J. (2024). *Collective AI on the Foundation AI: The pathway of digital transformation of intelligence (in press)*. CAFA Lab, Inc.

Choi, J., Kim, S., & Yoon, K. (2012-present). CAFA (v 2.0) Platform: Collective AI on the Foundation AI [Digital Platform]. CAFA Lab, Inc.

Drasgow, F., Luecht, R. M., & Bennett, R. E. (2006). *Technology and testing*. Educational measurement, 4, 471-515.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems, 35,* 22199-22213.

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. ETS Research Report Series, 1990(2), i-48.

Sheehan, K., & Lewis, C. (1992). Computerized mastery testing with nonequivalent testlets. *Applied Psychological Measurement, 16(1)*, 65-76.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. Journal of Educational measurement, 24(3), 185-201.

Wainer, H., Lewis, C., Kaplan, B., & Braswell, J. (1990). AN ADAPTIVE ALGEBRA TEST: A TESTLET-BASED, HIERARCHICALLY-STRUCTURED TEST WITH VALIDITY-BASED SCORING §. ETS Research Report Series, 1990(2), i-26.

S8-2: Paper Session - AI Topics 3

Chair: *Artur Pokropek*

Artur Pokropek, Marek Muszyński, & Tomasz Żółtak

Leveraging Approximate Areas of Interest (AAOIs) and Cursor Movement to Improve Validity in Low-Stakes Testing

Abstract: The shift to computer-based assessments brought new challenges and possibilities, such as collecting log data. Log data is a kind of paradata, additional, auxiliary information that is a by-product of any computerized assessment but can be used to gain additional insight, e.g., into testees' response processes and their level of motivation. While there is extensive research on utilizing paradata to enhance task design and comprehend student-task interactions, the multifaceted nature of paradata has often hindered substantial advancements. In many instances, the depth and richness of paradata are reduced merely to response time analysis. Only a few studies delved into the sequential structure of responses, but these investigations typically pertain to individual items. The rationale behind this limitation is that generalizing findings using the conventional approach to paradata becomes challenging. This is because tasks often possess distinct and non-comparable structures of states and actions, which hinders drawing consistent inferences across different tasks.

Our team has recently introduced a simplified methodology known as Approximate Areas of Interest (AAOIs) that offers a valuable alternative to the currently used data collection methods. This approach uses a flexible grid to identify distinct areas where respondents are most likely to interact with task elements on a screen. Using a vertical and horizontal splits system, the screen is divided into various Areas of Interest, such as instructions, items, response scale with labels, response area, and navigational buttons. The grid of AAOIs can be adjusted depending on the specific assessment design and screen layout. This adaptability allows researchers to tailor the grid to fit their study best, offering a practical and straightforward tool for understanding user engagement and assessment interaction patterns. This approach already yielded 95% accuracy in identifying disengaged respondents in the survey context (https://doi.org/10.1027/2151-2604/a000555).

In this presentation, we will show the result from a new study where 1,100 grade 7 and 8 students sit a cognitive test based on TIMSS math items. Students initially engage with only half of the test items. Upon completing this segment, they received notification that the primary assessment phase had concluded and that their scores would be based solely on their responses thus far. Subsequently, they were directed to guess the answers to the remaining items until they reached the end of the test. This approach of instructive guessing is rooted in methodologies adopted in several previous studies.

In tandem with the AAOIs framework, deep learning models were employed to detect responses from low motivation. Preliminary results show that the accuracy of rapid guessing detection is close to 95%, as in the survey context. The ability to detect low motivation in students holds the promise of significantly increasing the validity of low-stakes assessments. When unmotivated test-taking is identified and accounted for, the assessment data

reflects students' actual abilities more accurately than a blend of ability and varying motivation levels. This has profound implications for educational policymaking, curriculum design, and instructional interventions, as decisions will be based on more reliable data.

Burhanettin Ozdemir & Arwa Alsughayyer

Decision Tree-Based CAT in Clinical Settings: An Evaluation of Performance Metrics

Abstract: The purpose of this study is to investigate the feasibility of Decision Tree-based Computerized Adaptive Testing (DT-CAT) for clinical exams that utilize Standardized Objective Examination (SOE) and Objective Structured Clinical Examination (OSCE) stations that consist of polytomous items. Specifically, the study compares the performance of DT-CAT under various conditions, including different item bank sizes, item selection methods, and stopping rules, to determine the optimal DT-CAT algorithm and its impact on item exposure rate and ability estimates using the GRM models.

Simulations were conducted with item bank sizes of 100, 250, and 500 items, using both fixed and standard error based stopping rules. For the fixed stopping rules, each condition was assessed with averages of 6, 10, and 15 layers (items), and the randomesque method was applied with 10 random alternatives for each node at each level of the tree. Additionally, standard-error stopping rules included conditions such as stopping the exam when the standard error reached 0.20, 0.25, or 0.30. Two item selection methods, the Minimum Expected Posterior Variance (MEPV) and the Maximum Fisher Information (MFI), were employed. The performance was evaluated based on responses from 500 randomly generated examinees, with each test repeated 25 times per examinee to ensure the robustness of the results.

The performance of each condition was compared in terms of the percentage of items used, overall mean Standard Error of Measurement (SEM), and Root Mean Square Deviation (RMSD) of ability estimates. The results indicate that as the item bank size increases, the percentage of items used decreases significantly. For instance, with fixed stopping rules, the percentage of items used was 20% for the 100-item bank but dropped to 4.2% and 4.6% for the 500-item bank with MEPV and MFI item selection methods, respectively. Similarly, for the 250-item bank, the percentage of items used was 8.8%.

In terms of SEM, smaller item banks tend to have slightly higher SEM values compared to larger item banks. The RMSD of ability estimates generally decreased with increasing item bank sizes, indicating more accurate ability estimates with larger item banks. Notably, for the 500-item bank with 10 items and MEPV item selection, the RMSD was 0.414, compared to 0.571 for the 100-item bank with 10 items.

Comparing item selection methods, the MFI method tended to produce slightly higher SEM and RMSD values compared to MEPV, indicating less accurate ability estimates. For example, with 250 items and fixed stopping rules of 10 items, the RMSD for MFI was 0.455 compared to 0.426 for MEPV.

Recent advancements suggest that DTs offer significant advantages over traditional CATs. DT-CAT allows the complete test to be designed in advance, eliminating the computational cost associated with item selection during the test administration. Furthermore, DTs do not require the local independence condition of traditional CATs and can achieve accurate estimates of ability scores with fewer items. This study contributes to the growing body of evidence supporting the use of DT-CAT in clinical assessments, highlighting its efficiency and potential for broader application in adaptive testing environments.

Seong Il Kim

Learning Analytics on Learners' Dropout in a Language Learning Platform

Abstracts: Learning analytics (LA) involves the measurement, collection, analysis, and reporting of data about learners and their learning experiences. This research aims to explore the use of LA in predicting and understanding learners' dropout behavior in a childhood English learning platform in South Korea. The dataset includes extensive longitudinal records of learners' interactions with the platform, featuring over 200,000 records of content completion and time consumption, and similar volumes for AI-scored speaking test results.

The research employs a combination of survival analysis and deep learning techniques to achieve two primary objectives: predicting learner dropout and identifying the dominant factors influencing dropout/retention. The dataset underwent extensive preprocessing to derive relevant features such as time differences between user interactions and score distributions. The analysis also addressed issues related to data format transformations to retain as much information as possible from the raw data.

Initial survival analysis on 1-day dropout cases indicated that the frequency of repeated learning on the same content is statistically significant, but further refinement is needed for robust model performance. Extended analysis on 7-day and 30-day dropout cases revealed more pronounced impacts of variables such as 'Score_mean' and 'learned' on dropout events. Specifically, a decrease in the number of learned contents consistently reduced survival probability, and higher average scores were associated with a sharp decline in survival probability.

The study plans to employ both survival analysis and Recurrent Neural Networks (RNNs) to predict dropout, comparing the performance of these models. This dual approach aims to leverage the strengths of both traditional statistical methods and advanced machine learning techniques.

This research contributes to the field by providing insights into the factors influencing learner retention in educational platforms, and by developing predictive models that can inform interventions to reduce dropout rates. The findings can help educators and platform developers to better understand learner behavior and improve educational outcomes through targeted support and enhancements to the learning experience.

Table 1 Raw data sets and Variables

Data set1; Longitudinal Data, Record Length: 219,067

Variable(type)	Meta Info.	Number of unique values
ID(string)	user index	295
b_id(numeric)	content index	2278
stage(categorical)	content label	5
date(datetime)	Learning date and time	76080

Data set2; Longitudinal Data, Record Length: 217,375

Variable(type)	Meta Info.	Number of unique values
ID(string)	user index	285
b_id(numeric)	content index	1758
level(categorical)	content label	35
recording_rul(string)	URL to sound file	217375
score(numeric)	test results	217375

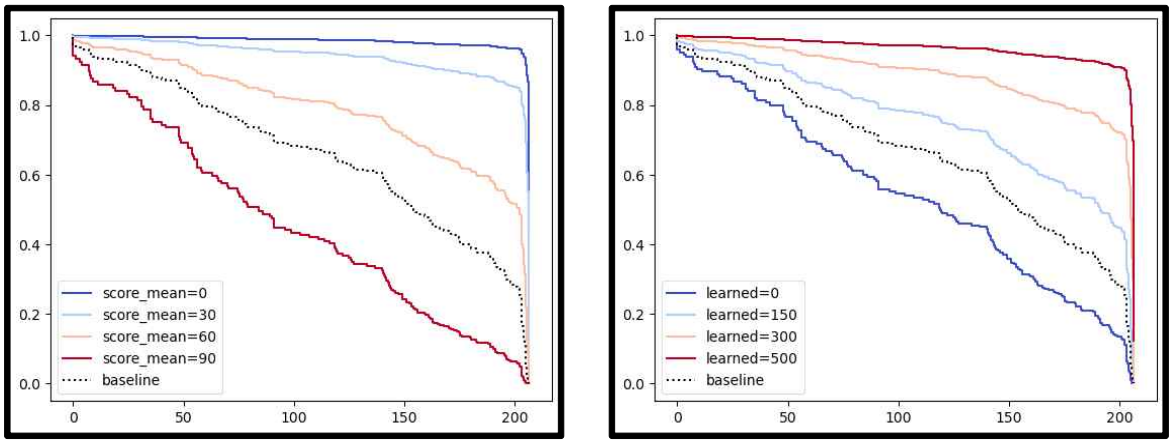


Figure 1 Significant variables on *7-days*+ dropout.

The variables 'score_mean' and 'learned' significantly affect the dropout event of 7 days or more. As the number of learned contents decreases, the survival probability consistently declines. Conversely, as the average score of the speaking evaluation increases, the survival probability also consistently declines

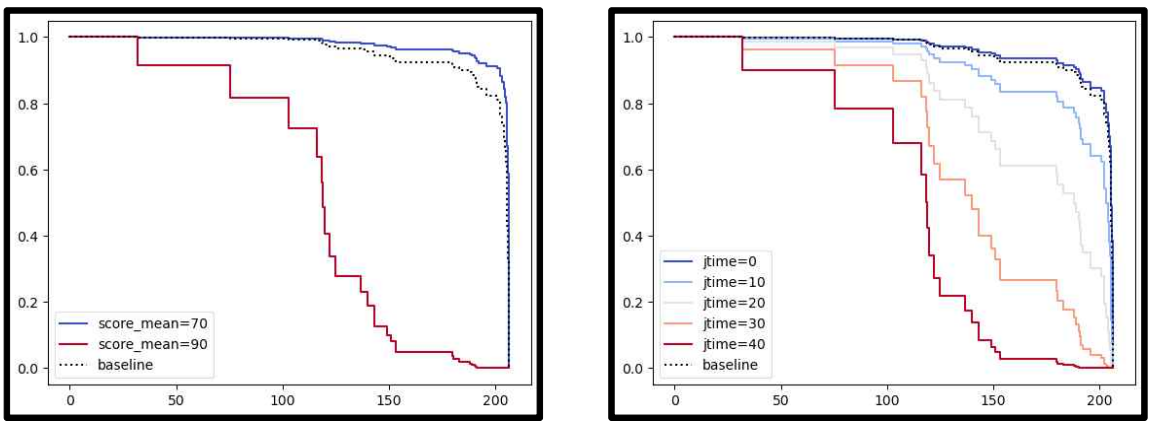


Figure 2 Total duration, *30-days*+ dropout.

The variables 'xcore_mean' and 'jtime' have a dramatic relationship with the dropout event of 30 days or more. Interestingly, a higher 'score_mean' is associated with a sharp decline in the survival probability.

As required, the qualification and source of the dataset are as follows:

1. Participated in the Data Analysis Contest of the Edutech Society (formerly the e-Learning Society).
2. Utilized the dataset provided by the Edutech Society (formerly the e-Learning Society) of an edutech company.

S8-3: Paper Session - Other Adaptive testing topics

Chair: *Semih Topuz*

Semih Topuz, Giray Berberoğlu, Kadriye Belgin Demirus, & Esra Kınay Çiçek

Can computer adaptive version of students' ratings of instruction provide valid results?

Abstract: Evaluation of instructional processes in universities has become a part of quality indicators in many countries. Most of the university administrations use the results of students' rating of instruction for promotion decisions as well. The validity of the results of students' ratings is an ongoing debate in academic literature. Some external factors might be interfering with the rating results, such as course load, grade leniency of the instructor, and past experiences of the instructor with students (d'Apollonia & Abrami, 1997; Greenwald & Gillmore, 1997; Griffin, 2004; Hoffman, 1978; Liaw & Goh, 2003; Marsh, 1982, 1983; Zhao & Gallant, 2012). Besides that, the reluctance of students to fill out the questionnaires and carelessness during the answering process is cited as important factors affecting the validity of the results (Bassett et al., 2017). In fact, students should spend a significant amount of time filling out the questionnaires for every course they take in a semester (Hoel & Dahl, 2019). It is anticipated that students may answer questionnaires carelessly since they have to answer the same questions for every course, which could negatively impact the validity of the results. Reducing the number of questions and using different sets of items for different courses could improve validity, as students are likely to pay more attention. The main goal of adaptive testing is to match item difficulty with the ability of the examinees. (Edwin Welch & Frick, 1993). If it is used in assessing affective measures such as students' rating of instruction, matching the item endorsement with the overall opinion of the students about the course becomes important. Deciding on how to start adaptive testing for individuals with different abilities is one of the major concerns of the researchers (Thompson & Weiss, 2011). In achievement testing, hybrid models with the same test form for all the examinees to estimate starting theta and continue the process to this estimation is a common approach used (Han, 2020; Wainer et al., 2000). This model may also be used in the students' ratings of instruction if there is information about students' overall opinions about the course before administering the items related to the effectiveness of instruction.

In the present study, computer adaptive testing of students' ratings of instruction will be simulated based on students' preconceived opinions about the course such as course difficulty and expected letter grades at the end of the semester. It is hypothesized that computer adaptive test administration will be more accurate and valid if students' overall opinions are used as a starting point in the adaptive test administration. This approach will also provide different sets of test items for different courses which might contribute to the accuracy and validity of the rating results. In general, in the present study, the effect of considering students' preconceived opinions about the course difficulty and expected letter grades will enhance the

adaptiveness of the whole process will be studied. This study will be a post hoc simulation research based on actual data.

References

Bassett, J., Cleveland, A., Acorn, D., Nix, M., & Snyder, T. (2017). Are they paying attention? Students' lack of motivation and attention potentially threaten the utility of course evaluations. *Assessment & Evaluation in Higher Education, 42*(3), 431-442. https://doi.org/10.1080/02602938.2015.1119801

d'Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist, 52*(11), 1198-1208. https://doi.org/10.1037/0003-066X.52.11.1198

Edwin Welch, R., & Frick, T. W. (1993). Computerized adaptive testing in instructional settings. *Educational Technology Research and Development, 41*(3), 47-62. https://doi.org/10.1007/BF02297357

Greenwald, A. G., & Gillmore, G. M. (1997). No pain, no gain? The importance of measuring course workload in student ratings of instruction. *Journal of Educational Psychology, 89*(4), 743-751. https://doi.org/10.1037/0022-0663.89.4.743

Griffin, B. W. (2004). Grading leniency, grade discrepancy, and student ratings of instruction. *Contemporary Educational Psychology, 29*(4), 410-425. https://doi.org/10.1016/j.cedpsych.2003.11.001

Han, K. (Chris) T. (2020). Framework for Developing Multistage Testing With Intersectional Routing for Short-Length Tests. *Applied Psychological Measurement, 44*(2), 87-102. https://doi.org/10.1177/0146621619837226

Hoel, A., & Dahl, T. I. (2019). Why bother? Student motivation to participate in student evaluations of teaching. *Assessment & Evaluation in Higher Education, 44*(3), 361-378. https://doi.org/10.1080/02602938.2018.1511969

Hoffman, R. G. (1978). Variables Affecting University Student Ratings of Instructor Behavior. *American Educational Research Journal, 15*(2), 287-299. https://doi.org/10.3102/00028312015002287

Liaw, S., & Goh, K. (2003). Evidence and control of biases in student evaluations of teaching. *International Journal of Educational Management, 17*(1), 37-43. https://doi.org/10.1108/09513540310456383

Marsh, H. W. (1982). Factors Affecting Students' Evaluations of the Same Course Taught by the Same Instructor on Different Occasions. *American Educational Research Journal, 19*(4), 485-497. https://doi.org/10.3102/00028312019004485

Marsh, H. W. (1983). Multitrait-Multimethod Analysis: Distinguishing between Items and Traits. *Educational and Psychological Measurement, 43*(2), 351-358. https://doi.org/10.1177/001316448304300204

Thompson, N. A., & Weiss, D. A. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research, and Evaluation, 16*(1), 1.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Routledge. https://doi.org/10.4324/9781410605931

Zhao, J., & Gallant, D. J. (2012). Student evaluation of instruction in higher education: Exploring issues of validity and reliability. *Assessment & Evaluation in Higher Education, 37*(2), 227-235. https://doi.org/10.1080/02602938.2010.523819

Luz Bay & Cassie Chen

Standard Setting Method for Computer Adaptive Tests

Abstract: One of the primary considerations in selecting a standard setting method is its compatibility with the design of the test itself. For a computer adaptive placement test the most often used standard setting method is Bookmarking. This presentation aims to compare Bookmarking with a standard setting method deemed more compatible with computer adaptive tests (CATs). Using data from a previously implemented standard setting, and implementation of a new method will be simulated for the purpose of results comparison.

The Computer Adaptive Standard Setting (CASS) method is without a doubt the method that is most consistent with CATs because it uses the actual CAT engine used for the test. Using an automotive analogy, thinking of the CAT test administration as a vehicle, say a sedan, the CASS method uses as the main tool another vehicle, say a coupe, that uses the exact same engine. This method was described on page 164 of Bay (2014), *State of the Art Standard Setting for State of the Art Assessments.*

> Applying the same rules and algorithms used to administer a computer adaptive test (CAT) to students, panelists would be presented with successive test items; however, instead of responding to the item, the panelists would consider the item and answer the following question: Will a student performing at the borderline of passing be able to respond to this question correctly? If a panelist's response is "yes," then a more difficult item would be presented next. If a panelist's response is "no," then a less difficult item would be presented next.
> …
> As the rating process continues, the algorithm would zero in on the item that represents the ability of a student at the borderline of passing. When this is determined with an acceptable degree of certainty (i.e., when the standard error is below a predesignated value), that ability estimate is the cut score set by that panelist. The panel cut score could then be determined by calculating the median or mean of the cut score set by the individual panelists.

The CASS is similar to a method referred to as "simulated student" method (SSM) which uses the actual CAT administration interface. The panelists' rating task for the SSM goes as follows:

> Based on your understanding of the BD respond to the CAT (using regular test administration)
> a. If the student matching the BD is able to answer the item correctly, the panelist will select the correct answer
> b. If the student matching the BD do not have the necessary KSA to respond to the item correctly, the panelist selects any of the wrong answers

The panelist's score at the termination of the CAT is their cut score. The median of the panelists' cut score is the panel cut score.

The main, and very important difference, between CASS and the SSM method is the interface used in item rating. To use CASS, a standard setting interface needs to be developed and replace the test administration interface.

Comparisons between CASS and SSM will focus on implementation issues, while numerical results will be the focus of the CASS comparison with Bookmarking. These comparisons can hopefully shed light on whether developing a standard setting interface to use over the ending of a CAT is an endeavor worth pursuing for publishers of computer adaptive tests.

Reference:

Bay, L. (2016). State of the art standard setting for state of the art assessment. In Lissitz, R. W., & Jiao, H. (Eds.). *The next generation of testing: Common core standards, Smarter-Balanced, PARCC, and the nationwide testing movement*. Charlotte: Information Age Publishing Inc.

Yoojin Chelsea Jang

Classting's AI diagnostics & personalized learning in Korean public education

The presentation outlines Classting's journey and innovations in personalized education through AI technology, particularly in South Korea. The company's mission is to make mastery learning accessible worldwide, leveraging AI-powered systems to support teachers in identifying student learning gaps and tailoring educational materials accordingly. Classting's AI solutions address two critical needs in public education: precise diagnostic assessments of student progress and the delivery of personalized learning materials.

At the core of Classting's technology are two AI systems. The first is the CLST Knowledge Tracing Engine, which predicts students' academic levels with high accuracy, especially in cold start environments. The second is Jello, a Retrieval-Augmented Generation (RAG) based large language model (LLM), designed to provide personalized tutoring to students and assist teachers in curriculum delivery. These technologies are powered by a massive dataset accumulated over 12 years, consisting of over 91 billion educational data points from Korean public education.

Classting's product suite enables personalized learning journeys through accurate diagnostics, effective learning recommendations, and goal-driven instruction, all aligned with international curriculum standards. The presentation emphasizes the importance of integrating both AI-powered diagnostics and personalized learning pathways to enhance students' cognitive and metacognitive skills, ultimately aiming to improve academic performance. Partnering with public institutions, Classting has demonstrated significant improvements in learning outcomes, positioning itself as a transformative force in education through AI.

S8-4: Paper Session - CAT applications 4

Chair: *Jeongwook Choi*

Jeongwook Choi, Sung-Soo Jung, Eun Kwang Choi, Kyung Sik Kim, Dong Gi Seo

Application of CAT to Comprehensive Clinical Medical Assessment in Korea Organizer

Abstract:

Purpose: The Korean Association of Medical Colleges (KAMC) is an association that aims to foster medical talent and advance medical education. The KAMC conducts a comprehensive clinical medical assessment twice a year, targeting all students across 40 domestic medical colleges. This assessment evaluates clinical medicine abilities from the first to the fourth grade of medical school and is conducted to compare abilities and track the growth of each university. The assessment has been administered utilizing computer-based testing. The assessment has transitioned to computerized adaptive testing to facilitate students in comparing their abilities with medical students nationwide and with their own previous performances. Additionally, transitioning to CAT provides more assessment opportunities for medical college students.

Method: To establish an item bank, items administered from 2019 to 2022 were calibrated using a two-parameter model. Items with appropriate item parameters(discrimination, difficulty) were retained, while those with inappropriate item parameters were excluded. Finally, a total of 1,528 items were used to construct the item bank. The comprehensive clinical medical CAT was administered through LIVECAT(CAT platform). The LIVECAT is an adaptive testing web platform based on a web browser. Test takers accessed the assessment via a web browser without the need for additional installation programs. The components of the administered CAT include the following: Theta Range: -5 ~ 5; Starting Rule: Randomly selected first five items; Item Selection Method: Maximum Fisher Information (MFI); Ability Estimation Method: Maximum Likelihood Estimation (MLE); Termination Rule: Standard Error of Estimation (SEE) <= 0.25 and Maximum Number of Items: 50. The CAT assessment was conducted every month from March 2023 to June 2023.
Result: During the evaluation period, 1,843 students took the assessment. The average ability score of the examinees was 0.5, and the highest ability score in each assessment was 5.0. Ability scores exhibited an approximate normal distribution. There was not much variation in the average and standard deviation of ability scores by each assessment. On average, 28 items were administered to each examinee, with 50 items given to students with low or high abilities.

Conclusion: For students with either low or high abilities, the assessment was terminated according to the maximum item rule. This occurred due to a mismatch between the abilities of these students and the difficulty distribution of the item bank. The item bank's difficulty follows a normal distribution with a mean of -0.5, resulting in a shortage of appropriate items for students with abilities below -1 or above 1.5. To assess students more accurately within these ability ranges, the development of items with difficulty levels

corresponding to these ranges is necessary. The evaluation was conducted without content balancing constraints. Due to the granularity of the assessment domains, there were insufficient items to ensure content balancing. For future Computerized Adaptive Testing (CAT) implementations that are more tailored to student abilities and assessment specifications, item development considering difficulty levels and domains is essential. Subsequent research on CAT with content balancing constraints will be required. The assessment implies the first official application of CAT in Korea.

Minjung Kim

Advancing GTELP English Tests with IRT: Improving Accuracy and Validity in Proficiency Assessment

Abstract: This study focuses on applying Item Response Theory (IRT) models to traditional English language assessments. The research encompasses several key areas essential for the transition from conventional testing methods to an IRT-based framework. Initially, we analyze and design test items to ensure they are suitable for IRT application, involving detailed evaluation of item difficulty, discrimination, and guessing parameters. We then select and apply appropriate IRT models, such as the 1-parameter, 2-parameter, and 3-parameter models, to assess their fit for English language tests. A pilot test is conducted to collect response data, which is analyzed to estimate item characteristics and examine the model's performance. The study also emphasizes the importance of verifying the reliability and validity of the IRT-based tests, ensuring they provide consistent and accurate assessments of language proficiency. We address potential challenges, such as avoiding bias and ensuring the test's fairness across diverse demographics. Finally, we refine and optimize the model, incorporating improvements based on data analysis, and explore the practical application of IRT in real-world educational settings. This research aims to enhance the precision and objectivity of English language assessments, providing a robust foundation for future applications of IRT in language testing.

Jeongin Cha, & Moonsoo Lee

Application of CAT to Adolescent Smartphone Overdependence scale

Abstract: With the advancement of technology, we are now able to utilize smartphones anytime and anywhere. However, behind this convenience lies the problem of overdependence. Particularly during adolescence, when self-regulation is challenging, it becomes difficult to curb smartphone usage. Therefore, schools and counseling settings are endeavoring to develop scales to prevent it.

The existing Adolescent Smartphone Addiction Scale comprises three factors: loss of control, salience, and problematic outcomes, which include physical health, mental health, interpersonal relationships, and productivity. Each factor consists of 4 items for loss of control and salience, and 3 items per subfactor for problematic outcomes, totaling 20 items. This scale uses a 4-point Likert scale, demonstrating high reliability with a Cronbach's alpha of .948. However, interviews with school teachers and counselors revealed concerns that the scale's 20 items were perceived as lengthy by students and raised issues regarding response sincerity. Particularly for screening and prevention purposes, identifying at-risk groups for smartphone overuse necessitates a more efficient approach, as students expressed difficulty in reading and accurately responding to more than 20 items. Therefore, this study aims to enhance efficiency while maintaining the accuracy of the assessment by applying Computerized Adaptive Testing (CAT) to the existing scale.

In the methodology, initially, the item parameters of the existing 20-item scale were analyzed. Correlation analysis indicated very high correlations (ranging from .75 to .88) between all items and their respective subfactors. Confirmatory factor analysis confirmed the appropriateness of the model based on CFI, TLI, and RMSEA values. Utilizing these results, item parameters for the four factors were analyzed using Item Response Theory (IRT), including analysis of Item Characteristic Curves (ICC) and Item Information Functions with Graded Response Models (GRM). Subsequently, simulations were conducted to create a new item pool approximately ten times larger than the original 20-item pool, based on existing item characteristics. A comprehensive Computerized Adaptive Testing (CAT) assessment was then developed using this item bank, exploring optimal item numbers and conditions for CAT development.

Through this study, efficiency can be enhanced while maintaining the accuracy of the scale. Implementing CAT allows the assessment originally consisting of 20 items to be conducted with approximately 10 items, simplifying administration without the need for pencil and paper, thereby increasing convenience in school and counseling settings. Furthermore, counselors and teachers noted that students with smartphone overuse find a 20-item assessment tedious, which can decrease response sincerity. By reducing the number of items, solutions to this issue can be proposed.

Conference Sponsors

Platinum Sponsors

연세대학교
YONSEI UNIVERSITY

Yonsei University is dedicated to educate future leaders of our society in the spirit of Christianity, fostering a strong and lasting commitment to the principles of truth and freedom. “If you hold to my teaching, you are really my disciples, then you will know the truth, and the truth will set you free” [John 8:31-32]. Yonsei University serves as the “alma mater” of all arts and sciences to nurture leaders who will contribute to the Korean and international society, in the ecumenical spirit of Christian teaching epitomized in its motto of “truth and freedom.” Yonseians will inherit mankind’s cultural legacy, and lead academic development through creativity and critical thinking. We will also serve others with an open heart, and contribute to the prosperity of humankind. With these missions, we Yonseians will demonstrate our leadership to realize the Yonsei spirit globally.

Gold Sponsors

CLASSTING

Classting (www.classting.com) is an EdTech AI company that developed "Classting AI," an AI-based Software as a Service (SaaS) platform. With the introduction of AI digital textbooks set to accelerate starting in March 2025, Classting AI leverages artificial intelligence to accurately assess individual students' levels and provide personalized learning paths. This approach transforms the traditional, one-size-fits-all public education model into a personalized system through AI technology, enabling an overall elevation of education standards.

Silver Sponsors

연세대학교
YONSEI UNIVERSITY

College of Educational Sciences, Yonsei University

Bronze Sponsors

ASC is dedicated to empowering organizations to leverage the benefits of modern psychometrics such as IRT and adaptive/multistage testing. Our cloud-based platform for item banking and online assessment makes it easy to develop and publish adaptive/multistage tests, whether for 100 or 100,000 examinees, with a wide range of security.

The Duolingo English Test is at the forefront of assessment science, integrating the latest AI to provide accurate and convenient results for anyone, anywhere. Our human-in-the-loop approach to AI allows us to prioritize the test taker experience (TTX) while ensuring the validity, accuracy, and security of every test session.

Pearson VUE has been a pioneer in the computer-based testing industry for decades, delivering close to 21 million certification and licensure exams annually in every industry from academia and admissions to IT and healthcare. We are the global leader in developing and delivering high-stakes exams via the world's most comprehensive network of nearly 20,000 highly secure test centers as well as online testing across more than 180 countries. Our leadership in the assessment industry is a result of our collaborative partnerships with a broad range of clients, from leading technology firms to government and regulatory agencies. For more information, please visit PearsonVUE.com.

CLASS-Analytics is designated to serve teachers and students as an innovative and cloud-based learning management system (LMS) focusing on administering linear and adaptive exams/tests. AI-based automatic item generation, adaptive testing system, and automatic essay scoring are main primary services.

College Board reaches more than 7 million students a year, helping them navigate the path from high school to college and career. Our not-for-profit membership organization was founded more than 120 years ago. We pioneered programs like the SAT® and AP® to expand opportunities for students and help them develop the skills they need. Our BigFuture® program helps students plan for college, pay for college, and explore careers.