



2026 IACAT Conference

Program book

10th Conference of the International Association for
Computerized Adaptive Testing

September 28-30, 2026

Summary

Welcome message	5
Welcome message from organizing committee	6
Organizing Committees	7
For your Orientation	8
Transportation	8
Conference Venue	10
Room Map	11
Events	12
Conference Dinner	12
Explore São Paulo	13
Schedule at a Glance	14
Abstracts	19
Workshop 1	19
Workshop 2	20
Workshop 3	21
Keynotes	22
Andreas Frey	22
Hua-Hua Chang	23
Maomi Ueno	24
Ricardo Primi	25
Wim J. van der Linden	26
S-01: Advancing Adaptive Measurement: Dr. David Weiss's IRT and CAT Lab Innovations in CAT Methodology and Adaptive Measures of Change	27
S-02: New IRT models to modeling different responses	31
S-03: Frontiers of Bayesian Adaptive Testing	34
S-04: Continuous Item Banking for a High-Stakes Certification Without Pretesting: An Operational Decomposition	39
S-05: The PROMIS, NIH Toolbox, and Mobile Toolbox Measurement Systems: Architecture, Psychometric Foundation, and the Challenges of Adaptive Health Measurement	44
S-06: EduCAT	48
Individual Presentations	49
Development and Evaluation of an IRT-Based Computerized Adaptive Testing System for General English Proficiency of College Students	49
Growth Norms for Computer Adaptive Tests: Evaluating Three Conditional Growth Percentile Methods	50
Enabling Computerized Adaptive Testing in large-scale educational assessments under limited connectivity conditions	52

Development of a Computerised Adaptive Test for Assessing Upper-Basic School Students' Mathematics-Ability in Nigeria: A Post-hoc Simulation Testing	53
Designing multi-stage CAT assessments of academic English language proficiency to give examinees the greatest opportunity to demonstrate that they meet performance level thresholds	54
Using Semantic Embeddings to Examine Content Overlap Between the IDCP-2 and PID-5	55
AI-Based Item Parameter Estimation as an Alternative to Pilot Testing in Computerized Adaptive Testing: A Simulation Study	56
Adaptive Testing Meets Contracting: Rethinking Success in the Age of AI	57
AI-Assisted Cognitive Pretesting for Personality Assessment Items: Comparing Profile-Conditioned Simulated Responses With Observed Human Data	58
Use of AI to Assess Metaphor Generation: Convergent And External Validity	59
Psychometrics Meets LLM: Diagnostic Assessment Copilot Design and Deployment	60
Automated Evaluation of Teacher Portfolios Using Large Language Models and Structured Educational Evidence Analysis	61
Methodology for the Empirical Generation and Validation of Performance Level Descriptors Using Generative Artificial Intelligence	63
A Cognitive Diagnosis Model for Latent Classification of Bounded Continuous Variables	64
Black-Box Variational Inference with REINFORCE for the DINA/DINO Model	65
Automatic Pre-Testing of Mathematics Assessment Items: Predicting Item Response Theory (IRT) Difficulty Parameter with Machine Learning Methods	66
Accelerating Item Piloting with NLP-Based Features and Bayesian CAT	68
LLM and Rule-Based Automatic Item Generation: A Two-Study Investigation in Large-Scale Literacy Assessment	69
An IRT-Base Adaptive Measure Of Conscientiousness In The Big Five Framework	71
Adaptive Test of Vocational Interests (TIPA): Development and psychometric properties of a computerized assessment measure	72
Using Item Response Theory to Develop a Parsimonious Measure of Psychological Need Satisfaction at Work: Implications for Decent Work Research and Social Justice	73
Review and Requalification of a Workforce Competency Assessment Tool for Technical Education	74
Design and Implementation of Practice-Oriented, Large-Scale Certification Exams: The Case of ANBIMA's Distribution Certifications	75
Form Assembly: An AI Agent Workflow for Building Parallel Spring Forms for a Diagnostic Classification Assessment	76
A Stochastic Constrained Test Assembly Method Applied to Multistage Adaptive Testing	77
Designing multi-stage CAT assessments of academic English language proficiency to give examinees the greatest opportunity to demonstrate that they meet performance level thresholds	78
parATA: an R-package for Automated Test Assembly	79
Beyond Correct and Incorrect: A Nested Logit Computerized Adaptive Test Using Distractor Information to Assess Reading Proficiency	80

Computerized Adaptive Testing Using Nonparametric Item Response Theory Models	81
Addressing Misclassification Bias in Computerized Adaptive Testing with Latent Performance Standards	82
Testlet-Based Termination in Relaxed CAT: Reducing Assessment Length and Testlet Exposure	83
Computer Adaptive Testing with small item bank: Improved precision and reduced opportunities for cheating	85
Posters	86
Applications and Implications of Unidimensional, Multidimensional, and Bifactor Models in Measuring Students' Socioeconomic Status	86
Can Large Language Models Predict 3PL Item Parameters in Brazilian Portuguese? An Exploratory Study with ENEM Items	87
CAT API: Exploring the Application of Computerized Adaptive Testing for Formative and Summative Assessments	88
Comparing Large Language Models for Psychological Scale Translation: Translation Quality Scores and Blinded Item Selection in the Brazilian Adaptation of the Teacher Autonomous Behavior Scale	89
Developing a German reading comprehension CAT for learning progress assessment	90
Psychometric Calibration of the Adolescent Externalizing Behaviors Bank (AEBank) Using the Graded Response Model: Item Functioning, Information, and Differential Item Functioning by Sex	91
Psychometric Calibration of the CIMEC-40 for Eating Disorder Risk Assessment in Adults: Development of a Screening Tool	92
Psychometric Evaluation of a Social Competence Item Bank for Children: A Rasch Partial Credit Model and CAT Simulation Approach	93
The effect of (mis)judged provisional item difficulties on calibrations: Evidence from online pretesting in Linguaskill and English Skills Test	94
Organizing Institutions	95
Supporting Institutions	97
Conference Sponsors	98

Welcome message from organizing committee

It is with great pleasure that we welcome you to IACAT 2026 and to São Paulo, as the conference comes to Brazil — and to Latin America — for the first time.

We are pleased to host IACAT 2026 and are committed to a program that combines the scientific rigor traditionally associated with IACAT with the social and cultural richness of Brazil.

The conference will be held at an exceptional venue on Avenida Paulista, one of São Paulo's main cultural and touristic centers, offering excellent infrastructure in a vibrant urban setting.

The scientific program will follow the well-established structure of previous IACAT conferences, including workshops, keynotes, symposia, and regular sessions, with the addition of a poster session introduced in this edition.

For participants interested in exploring the city and beyond, we have partnered with a tourism agency to offer optional guided experiences. These include a curated walking tour along Avenida Paulista, highlighting its cultural and gastronomic landmarks, as well as an optional post-conference visit to the beautiful city of Rio de Janeiro.

Further details will be shared on this website as they become available. Should you have any questions, please feel free to contact us by email — we will be happy to assist you.

We very much look forward to welcoming you to IACAT 2026 and to a rewarding experience in Brazil.

Warm regards,

Mariana Curi

Organizing Committees

Local

- Mariana Curi (University of São Paulo, Brazil) – chair
- Thiago Costa (Municipal Department of Education of the city of São Paulo, Brazil)
- Karina Stefanin (SESI, Brazil)
- Caio Naberezny Azevedo (University of Campinas, Brazil)
- Girlene Ribeiro de Jesus (Federal University of Brasília, Brazil)

IACAT Board

- Chun Wang (Incoming IACAT president)
- Kathi Gialucca (IACAT Treasurer)
- Nathan Thompson (IACAT Executive Director)

Student Assistants

- Antonina Curi Buscaglia (Student Volunteer)
- Felli Costa Pazzotto (University of São Paulo, Brazil)
- Guilherme Gonçalves Rodrigues (University of São Paulo, Brazil)
- João Victor Silva Guerra (University of São Paulo, Brazil)
- Maria Fernanda Rodrigues Perosa (University of São Paulo, Brazil)
- Vitor Hugo Boraschi (University of São Paulo, Brazil)

Scientific Committee

- Caio L. Naberezny Azevedo (University of Campinas, Brazil)
- Jorge Luiz Bazán (University of São Paulo, Brazil)
- Mariana Curi (University of São Paulo, Brazil)
- Dalton Francisco de Andrade (Federal University of Santa Catarina, Brazil)
- Girlene Ribeiro de Jesus (Federal University of Brasília, Brazil)
- Ricardo Primi (São Francisco University, Brazil)
- Héilton Ribeiro Tavares (Federal University of Pará, Brazil)
- Chun Wang (Incoming IACAT president)
- Wim van der Linden (IACAT president)

For your Orientation

Transportation

As a general frame of reference, 1BRL = 0.19 USD or 0.16 Euro as of January 30, 2026.

By Train

Take the Airport Express train from the Airport Station to Luz Station. Transfer to Metro Line 2-Green and get off at Trianon-Masp station, right next to the Centro Cultural FIESP on Paulista Av. (USD \$1-\$3)



By Executive Bus

Airport Bus Service is an airport shuttle bus that goes from the airport terminals to key points in São Paulo, including Paulista Avenue. Remember to check the exact stop list when buying a ticket. (USD \$7-\$10)

[Check them out](#)

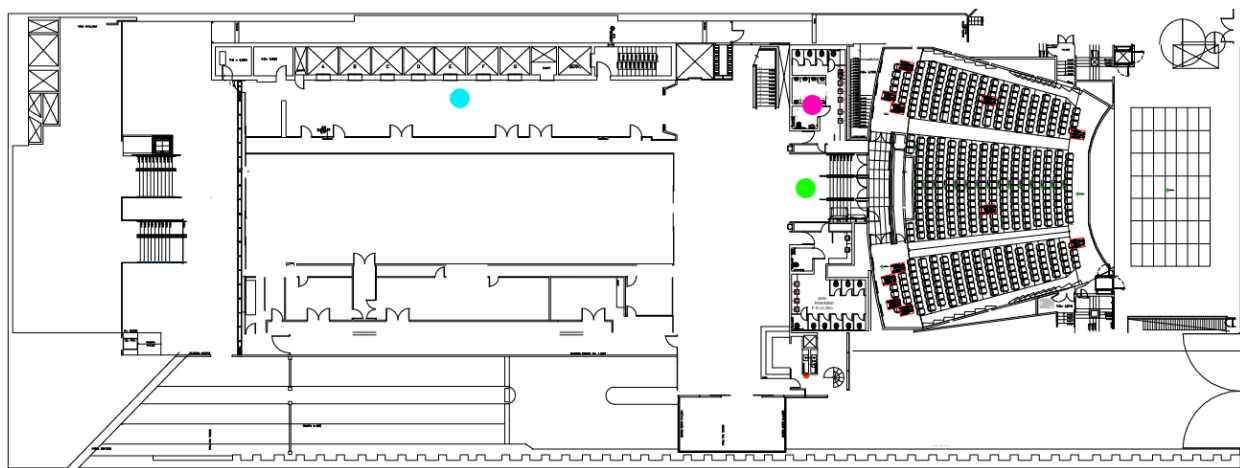
By Taxi

At GRU, there are official taxi services that offer fixed prices from the airport to destinations in São Paulo, including Paulista Avenue. Terminals 1,2, and 3 all have taxi counters/stands right outside the arrivals area. (USD \$20 - \$30)

By Uber

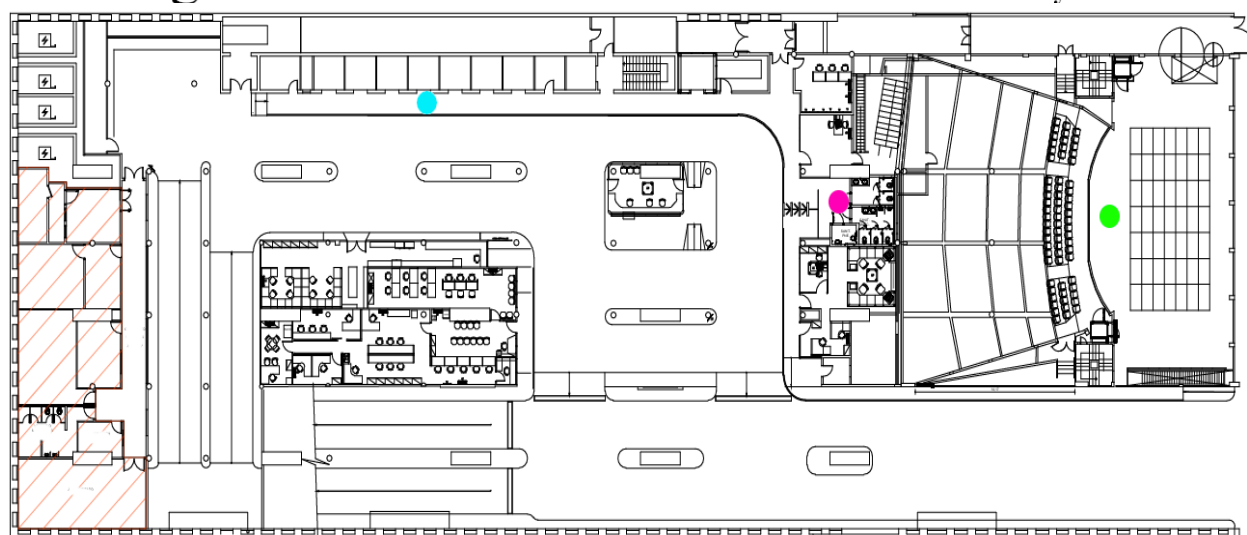
You can request an Uber ride in the app after you land and collect your luggage. The driver will pick you up at an official rideshare pickup area at the airport. Each terminal has a designated rideshare pickup zone usually next to the departures or car parking area designated by Uber. (USD \$15 - \$35)

Conference Venue



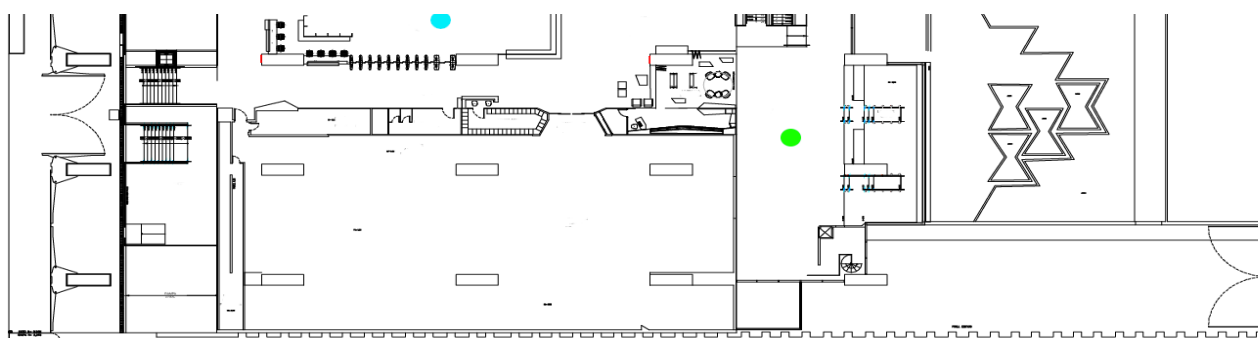
Lower ground floor

● Elevator Hall
● Foyer Theater



1st Basement level

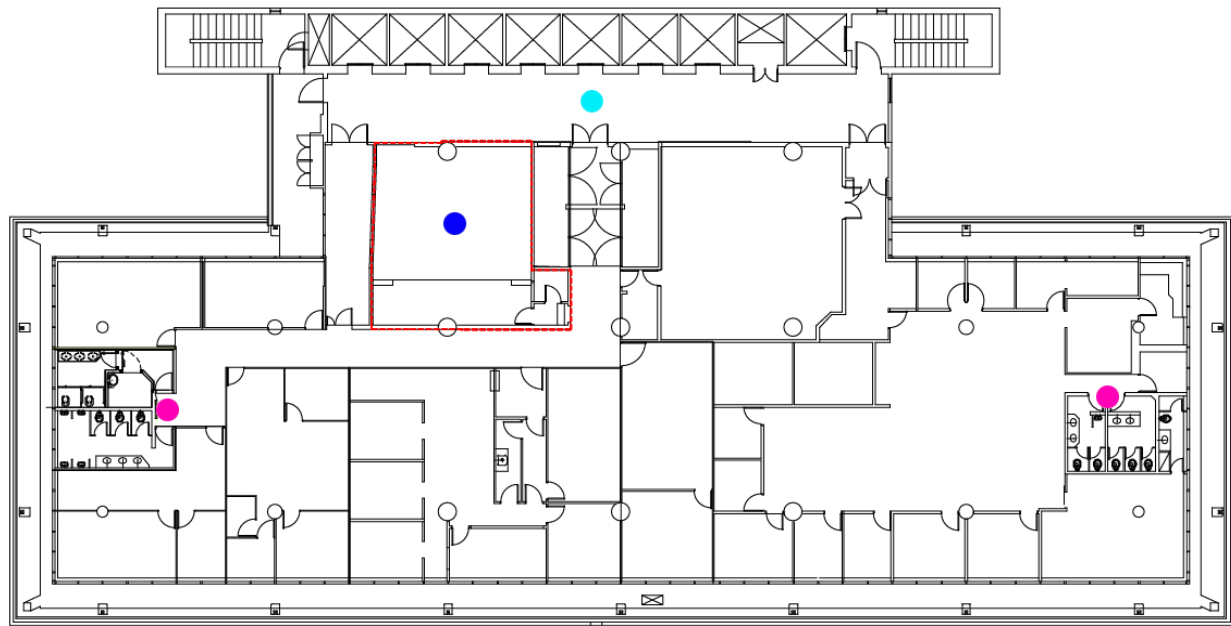
● Elevator Hall
● Foyer
● Bathroom



Upper ground floor

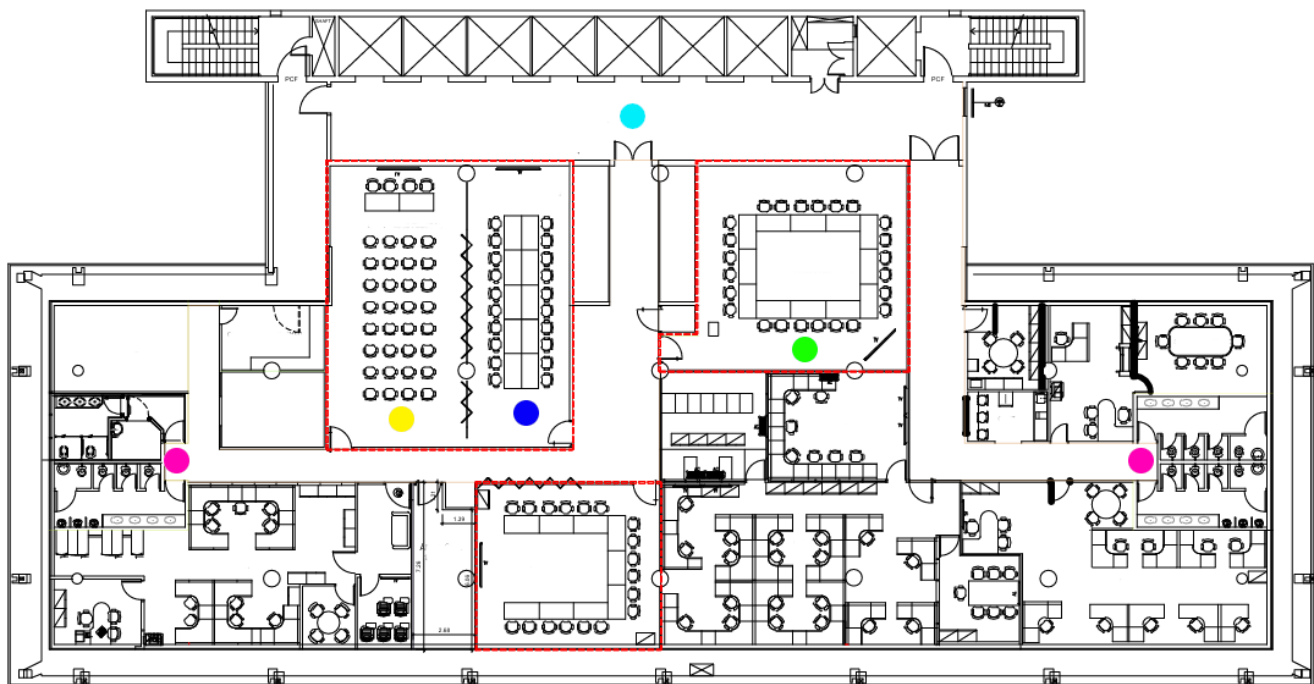
● Elevator Hall
● Foyer

Room Map



10th floor

- Bathroom
- Elevator Hall
- Auditorium



11th floor

- Room 1110
- Room 1120
- Room 1150
- Bathroom
- Elevator Hall

Events

Welcome Reception

On Tuesday, September 28, 6:00 PM - 8:00 PM we look forward to welcoming you to a reception with cocktail-style reception. This reception is complimentary for all registered attendees.

The Welcome Reception will feature a 20-minute cultural performance, scheduled for around 7PM

Conference Dinner

The Conference Dinner will take place at Catedral da Sé, a Neo-Gothic cathedral located at Praça da Sé. The program begins on Tuesday (29) at 4:30 PM, starting with a guided tour of the cathedral. Its construction began in 1913 and the cathedral was inaugurated in 1954. It is one of the largest Neo-Gothic cathedrals in the world, with seating capacity for 900 people. The evening continues with a concert performed by the Coro Luther King, followed by dinner in the cathedral halls organized by the renowned chef Gil Gondim. The funds raised through these events are allocated to support two social projects — Missão Belém, which provides assistance to people in vulnerable situations, and Marco Zero da Gastronomia, which offers professional training to the same population — as well as to the maintenance of the cathedral. Round-trip transportation from the conference venue on Avenida Paulista is provided. 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.

Explore São Paulo

The main attractions around Avenida Paulista, as well as the Sé Cathedral, are already included in activities organized by the conference. For those who would like to explore more of São Paulo, some of the city's additional well-known attractions are listed below:

- Ibirapuera Park – one of the city's most famous parks, with green spaces, museums, and cultural venues.
- Museu do Ipiranga – a historic museum dedicated to Brazilian history, set in a beautiful park.
- Liberdade – a lively neighborhood known for its Japanese and Asian cultural heritage, restaurants, and shops.
- Vila Madalena – a lively neighborhood known for its street art, bars, restaurants, cafés, and vibrant nightlife.
- Pinacoteca de São Paulo – one of Brazil's leading art museums, with an important collection of Brazilian art.
- Luz Station & Museum of the Portuguese Language – a historic railway station and a museum dedicated to the Portuguese language and its cultural heritage.
- Municipal Market (Mercadão) – a traditional market known for its variety of fresh produce and Brazilian foods.
- Historic Center – Sé Cathedral, Pátio do Colégio, São Bento Monastery, Municipal Theatre, Farol Santander, Martinelli Building, Copan Building, Municipal Market (Mercadão), Praça da República, Edifício Matarazzo
- University of São Paulo (USP) – one of Brazil's leading universities, with a large campus and several museums and cultural spaces.

For more information about attractions, culture, gastronomy, tours, and different areas of the city, visit the official website of the [São Paulo Municipal Tourism Office](#)

Schedule at a Glance

The conference consists of workshops, keynote lectures, invited symposia, a submitted symposium and thematic paper sessions. The coffee breaks during the conference and the welcome reception on day 1 are covered by your registration fee. The Schedule at a glance on the next pages gives an overview over the IACAT 2026 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 1 - 28/9 (Monday)		
Time	Activity	Venue
08:00 - 10:00	Workshop 1	Room 1110
	Workshop 2	Auditorium 10th floor
	Workshop 3	Room 1150
10:00 - 10:30	COFFEE BREAK	Room 1120
10:30 - 12:00	Workshop 1	Room 1110
	Workshop 2	Auditorium 10th floor
	Workshop 3	Room 1150
12:00 - 14:00	LUNCH	Free
14:00 - 14:30	Openning ceremony	Theater
14:30 - 14:45	Introduction of board members and Welcome to incoming president Chun Wang	Theater
14:45 - 15:45	Keynote 1 - Wim van der Linden (Presidential address) Simplifying large-scale educational assessments through Bayesian adaptive testing Chair: Chun Wang	Theater
15:45 - 16:15	COFFEE BREAK	Foyer Theater
16:15 - 17:15	Keynote 2 - Andreas Frey Selecting Empirical Prior Distributions for the Ability Parameters in Large-Scale Educational Assessments chair: Jonathan Templin	Theater
17:15 - 18:00	Awards and Travel Grants chair: Nathan Thompson	Theater
18:00 - 19:00	Poster session (9 works)	Foyer Theater
19:00	Welcome Reception	Foyer Theater

DAY 2 - 29/9 (Tuesday)

Time	Activity	Venue
08:00 - 9:40	Symposia: EduCAT - chair: Dalton F. Andrade	Theater
	Adaptive Assessment in Practice: Applications of CAT at SESI and in the ANPAD Test (Franciele Sena)	
	Item Calibration with Avatars: Innovation and Equity in Assessment (Dalton F. De Andrade)	
	Individual presentations: Applications of CAT (Education and Language) – chair: Luis Valdivieso	Room 1150
	Development and Evaluation of an IRT-Based CAT System for General English Proficiency of College Students (Dong Seo)	
	Growth Norms for Computer Adaptive Tests: Evaluating Three Conditional Growth Percentile Methods (Luciana Cancado)	
	Enabling Computerized Adaptive Testing in large-scale educational assessments under limited connectivity conditions (Gustavo Silva)	
	Development of a Computerised Adaptive Test for Assessing Upper-Basic School Students? Mathematics-Ability in Nigeria: A Post-hoc Simulation Testing (ISAAC OLAWALE IFINJU)	
	Designing multi-stage CAT assessments of academic English language proficiency to give examinees the greatest opportunity to demonstrate that they meet performance level thresholds. (Stephen Walker)	
	Individual presentations: Other AI Topics in Psychometrics – chair: Joaquim Soares Neto	Room 1110
	Using Semantic Embeddings to Examine Content Overlap Between the IDCP-2 and PID-5 (Pedro Godoy dos Santos)	
	AI-Based Item Parameter Estimation as an Alternative to Pilot Testing in CAT: A Simulation Study (Zeus Bellido)	
	Adaptive Testing Meets Contracting: Rethinking Success in the Age of AI (Xiangen Hu)	
	AI-Assisted Cognitive Pretesting for Personality Assessment Items: Comparing Profile-Conditioned Simulated Responses With Observed Human Data (Diego Forteza)	
	Use of AI to Assess Metaphor Generation: Convergent And External Validity (Beatriz Chrispim)	
9:40 - 10:00	COFFEE BREAK	Foyer Theater
10:00 - 11:00	Keynote 3: Hua-hua Chang – chair: Wim van der Linden Positioning Computerized Adaptive Testing as a Foundation for Personalized Learning	Theater
11:00 - 12:00	Keynote 4: Maomi Ueno – chair: Wim van der Linden High-Stakes Assessments with Process-Integrated IRT and AI: From Empirical Evidence to Computational Design Strategies for Computerized Adaptive Testing	Theater
12:00 - 14:00	LUNCH	Free
13:00 - 14:00	FIESP Cultural Center guided tour	
14:00 - 15:40	Individual presentations: Other AI Topics in Psychometrics – chair: Richard Gershon	Room 1150
	Psychometrics Meets LLM: Diagnostic Assessment Copilot Design and Deployment (Chun Wang)	
	Automated Evaluation of Teacher Portfolios Using LLM and Structured Educational Evidence Analysis (Gustavo Silva)	
	Methodology for the Empirical Generation and Validation of Performance Level Descriptors Using Generative AI (Daniela Jimenez)	
	Symposia: New IRT models to modeling different responses – chair: Jorge Bazan	Auditorium 10th floor

	Revisiting the Skew-Normal IRT family: Bayesian estimation and application (Cristian Bayes)	
	A robust GNBk count item response model (Luis Valdivieso)	
	Bayesian Estimation for a class of ability-based guessing IRT models (Alex de la Cruz Huayanay)	
	Symposia: Frontiers of Bayesian Adaptive Testing – chair: Jonathan Templin	
	Prior Specification for Ability Estimation in Bayesian Adaptive Testing (Jonathan Templin)	
	Spectral Efficiency in Multidimensional Bayesian Adaptive Testing: Comparing Fully Bayesian, Sequential Two-Level, and Globally Optimized Two-Level Frameworks (Seung Choi)	
	Bayesian Adaptive Testing for the Answer-Until-Correct-Response Format (Andreas Frey)	
	Bayesian Optimal Large-Scale CAT Using Constraint-Governed Augmentation (Richard Patz)	Theater
15:40 - 16:10	COFFEE BREAK	Foyer Theater
16:30 – 17:30	FIESP Cultural Center guided tour	
16:30 - 22:00	Departure for the Cathedral and Conference Dinner	Free

DAY 3 - 30/9 (Wednesday)

Horário	Atividade	Venue
08:00 - 09:00	Early Career Award: Dr Hyeon-Ah Kang – chair: Nathan Thompson Beyond Information: Balancing Measurement Precision, Testing Efficiency, and Item Security in Technology-Enhanced Adaptive Testing	Theater
09:15 - 10:30	Individual presentations: Cognitive diagnostic and IRT models - chair: Hua-hua Chang	Room 1110
	A Cognitive Diagnosis Model for Latent Classification of Bounded Continuous Variables (Eduardo Schneider Bueno de Oliveira)	
	Black-Box Variational Inference with REINFORCE for the DINA/DINO Model (Flavio Barros)	
	Symposia: Continuous Item Banking for a High-Stakes Certification Without Pretesting: An Operational Decomposition chair: Andrea Burgos de A. Mangabeira	Theater
	Interactive Branching-Scenario Items in High-Stakes Financial Certification: Bundle Scoring Without Venue Dependence (Andrea Burgos de Azevedo Mangabeira)	
	Closing the Loop on Item Demand: A Decomposable Formalism for Continuous Item Banking (Marco Pepe)	
	Continuous Calibration of an Operational Item Bank Without Pretesting (Erica Ruiz)	
	Individual Presentations: Item generation, item banking, form assembly, and others chair: Alexandre Jaloto	Room 1150
	Form Assembly: An AI Agent Workflow for Building Parallel Spring Forms for a Diagnostic Classification Assessment (Ann Hu)	
	Accelerating Item Piloting with NLP-Based Features and Bayesian CAT (Steven Nydick) ¹	
	LLM and Rule-Based Automatic Item Generation: A Two-Study Investigation in Large-Scale Literacy Assessment (Lucas Larcher)	
	COFFEE BREAK	Room 1120 Foyer Theater
10:00 - 10:30	Symposia: The PROMIS, NIH Toolbox, and Mobile Toolbox Measurement Systems: Architecture, Psychometric Foundation, and the Challenges of Adaptive Health Measurement – chair: Richard Gershon	Theater
10:45 - 12:00	Computerized Adaptive Testing (CAT) for Patient-Reported Outcomes Measurement Information System (PROMIS): Evaluation of Stopping Rules for the Pediatric Item Banks (Jiwon Kim)	
	NIH Toolbox Calibration and Scaling for a multi-stage episodic memory test across the lifespan (Emily Ho)	
	CAT on Mobile Devices: Assessments of Verbal Ability for Remote Administration in Mobile Toolbox (Y. Catherine Han)	
	Individual Presentations: Applications of CAT – chair: Mariana Curi	Room 1110
	An IRT-Based Adaptive Measure Of Conscientiousness In The Big Five Framework (Gabriela S. Lozzia)	
	Adaptive Test of Vocational Interests (TIPA): Development and psychometric properties of a computerized assessment measure (Gustavo Martins)	
	Using Item Response Theory to Develop a Parsimonious Measure of Psychological Need Satisfaction at Work (Alexsandro Luiz de Andrade)	
	Individual Presentations: Applications of CAT – chair: Roberta Palacios	Room 1150
	Review and Requalification of a Workforce Competency Assessment Tool for Technical Education (Paloma de Lima Santos)	

	Design and Implementation of Practice-Oriented, Large-Scale Certification Exams: The Case of ANBIMA's Distribution Certifications (Andrea Burgos de Azevedo Mangabeira)	
12:00 - 14:00	LUNCH	Free
14:00 - 15:40	Symposia: Advancing Adaptive Measurement: Dr. David Weiss's IRT and CAT Lab Innovations in CAT Methodology and Adaptive Measures of Change – chair: Robert Chapman	Auditorium 10th floor
	Change Pattern Detection in Adaptive Measure of Individual Change (Raj Wahlquist)	
	Interactions Between Termination Criteria and Ability Estimators in Computerized Adaptive Testing (Xinyu Liu)	
	One of these things is not like the other: Setting Thresholds for Equivalence of Item Response Theory Parameter Estimation (Advancing Adaptive Measurement) (Robert Chapman)	
	Tracing the Roots and Branching into the Future: The Academic Legacy of Dr. David J. Weiss (Advancing Adaptive Measurement) (Robert Chapman)	
	Individual Presentations: MST, Item generation, item banking, form assembly, and others – chair: Heliton R. Tavares	Theater
	Automatic Pre-Testing of Mathematics Assessment Items: Predicting Item Response Theory (IRT) Difficulty Parameter with Machine Learning Methods (Patrick Canto de Carvalho)	
	A Stochastic Constrained Test Assembly Method Applied to Multistage Adaptive Testing (Alina von Davier)	
	Designing multi-stage CAT assessments of academic English language proficiency to give examinees the greatest opportunity to demonstrate that they meet performance level thresholds (Stephen Walker)	
	parATA: an R-package for Automated Test Assembly (Angela Verschoor)	
	Individual Presentations: Statistical methodology – chair: Ricardo Primi	Room 1150
	Beyond Correct and Incorrect: A Nested Logit Computerized Adaptive Test Using Distractor Information to Assess Reading Proficiency (Thiago Costa)	
	Computerized Adaptive Testing Using Nonparametric Item Response Theory Models (Cecilia Marconi)	
	Addressing Misclassification Bias in Computerized Adaptive Testing with Latent Performance Standards (Bartosz Kondrtek)	
	Testlet-Based Termination in Relaxed CAT: Reducing Assessment Length and Testlet Exposure (Luciana Cancado)	
	Computer Adaptive Testing with small item bank: Improved precision and reduced opportunities for cheating (Rense Lange)	
15:40 - 16:10	COFFEE BREAK	Foyer Teatro
16:10 - 17:10	Keynote 5: Ricardo Primi Navigating the Psychometric AI Revolution: Innovations in Scoring and Validity chair: Mariana Curi	Teatro
17:10 - 17:30	Ending Ceremony	

Abstracts

Workshop 1

Introduction to IRT and CAT

Nathan Thompson (ASC)

New to CAT? This workshop is for you. We will start with an introduction to the CAT algorithm, including item selection, exposure constraints, scoring, and termination rules. We will discuss how this algorithm can be adapted to various approaches, including multistage testing, and how to utilize simulations to design and validate your CAT. Workshop assumes you are familiar with item response theory, and ready to apply it in adaptive testing.

Workshop 2

AI Applications in Adaptive Testing

Duanli Yan (ETS)

Alina A. von Davier (Duolingo)

In the era of artificial intelligence (AI), the field of educational testing faces significant challenges, particularly in test development and scoring in adaptive assessment, two key innovations include automated item generation (AIG) and automated scoring (AS). Only recently that generative AI has facilitated the development of complex test items on a large scale. We introduce “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 present an overview of the latest natural language processing (NLP) techniques and large language models for AIG, alongside psychometric principles and practices for test development. We discuss the application of engineering principles in designing efficient item production processes (Luecht, 2008; Dede et al, 2018; von Davier, 2017). As AS becomes an integral part of the assessment landscape due to their advantages in reporting time, cost, objectivity, consistency, transparency, and feedback. We aim to demystify AS and provide a comprehensive understanding of its workings. We offer an overview of the design, development, evaluation, and quality control of automated scoring systems, along with practical advice and considerations for practitioners on the applications of these systems into formative and summative assessments (Yan, Rupp, & Foltz, 2020). We will share the most recent AI applications in educational learning and assessment based on our upcoming volume *AI for Measurement in Educational Learning and Assessment* (von Davier and Yan, 2026).

Workshop 3

The Shadow-Test Approach to Adaptive Testing

Seung W. Choi (UT-Austin)

Richard J. Patz (University of California-Berkeley)

Wim J. van der Linden (University of Twente)

It is tempting to think of adaptive testing as a more complicated version of the problem of automated assembly of fixed test forms. But thanks to a simple twist of the latter, the problem is easier to solve, always produces maximum information about the ability of each of the test takers, meets any blueprint in force for the test, and easily generalizes to testing formats with varying degrees of adaptation such as item-level adaptive testing, linear-on-the-fly testing, standard multistage adaptive testing, and multistage testing with adaptive routing tests.

The course has three different parts. In the first part, 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 parts are to introduce two software packages available for the implementation of the shadow-test approach to adaptive testing, offering the participants hands-on experience with the R package *TestDesign* and *Optimal CAT*, a currently freely downloadable microservice available for easy integration with common test delivery systems.

Participants are expected to bring their own laptops to the course. Handouts will be sent to the participants prior to the conference to prepare for the course and review its content afterwards.

Keynotes

Andreas Frey

Andreas Frey is Professor for Educational Psychology with focus on Measurement, Evaluation & Counselling, managing director of the department for Educational Psychology and scientific director of the counseling center MAINKIND, all at Goethe University Frankfurt, Germany.

Previously, he was Professor for Empirical Methods in Educational Research at Friedrich Schiller University Jena, and Professor II for Educational Measurement at the University of Oslo, Norway. He was responsible for the national IRT scaling of PISA 2006 in Germany and has since then been involved in various national and international large-scale assessments of student achievement. Currently, he is member of the German PIRLS 2026 consortium. the Andreas is Associate Editor of the journal Behavior Research Methods and has several other editorial roles such as Consulting Editor of the Journal of Computerized Adaptive Testing. He serves as European Representative in the IACAT board and as President of the section Methods & Evaluation of the German Psychological Association (DGPs). His research interests are item response theory, computerized adaptive testing, methods of large-scale assessments, and theory-based and responsible use of AI for measurement purposes. His work was published in more than 140 papers in peer-reviewed journals and book chapters.

The title of his keynote is *Selecting empirical prior distributions for the ability parameters in large-scale educational assessments*. (Co-author Aron Fink)

Hua-Hua Chang

Hua-Hua Chang is a Professor at Purdue University in the United States.

Dr. Chang's research spans both theoretical and applied domains, including Computerized Adaptive Testing (CAT), Cognitive Diagnosis, and Differential Item Functioning. Dr. Chang is also at the forefront of developing innovative web-based assessment tools to enhance personalized learning. Dr. Chang is widely recognized for his pioneering contributions to CAT, where his innovative algorithms have significantly advanced the delivery of personalized assessments. More recently, his research has explored the intersection of CAT and generative AI, demonstrating how adaptive testing data can inform AI systems to produce learner feedback that is both concise and richly detailed. Dr. Chang's work has earned him accolades from leading professional organizations, including AERA, NCME, and APA. He is also honored as a fellow of both the American Educational Research Association (AERA) and the American Statistical Association (ASA).

Maomi Ueno

Maomi Ueno is a Professor at the Graduate School of Informatics and engineering at The University of Electro-Communications in Japan.

He received his B.S. and M.S. degrees in Education from Kobe University in 1989 and 1992, respectively, and his Ph.D. in Computer Science from Tokyo Institute of Technology in 1994. He was an assistant professor at the Graduate School of System Science at Tokyo Institute of Technology from 1994 to 1996. He is now a professor at the Graduate School of Informatics and Engineering at The University of Electro-Communications. His research interests span a range of areas, including data science, Bayesian statistics, and algorithm development, resulting in over 150 publications in international journals and conferences. He has been active in both local and international academic societies, including the program committees of AAAI, IJCAI, PGM, AISTAT, AIED, and EDM. He has also served as Local Steering Committee Chair for ICALT 2007 and AIED 2023. He received the Best Paper Award at the 20th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2008), as well as Outstanding Paper Awards at e-Learn 2004, e-Learn 2005, e-Learn 2007, and ED-MEDIA 2008.

The title of his keynote is *Addressing the Accuracy-Exposure Trade-off: Computer Science and AI approach for Computerized Adaptive Testing*.

Ricardo Primi

Ricardo Primi is a Psychologist, Associate Professor at Universidade São Francisco (USF), Campinas, Brazil.

He received his Ph.D. in School Psychology and Human Development at University of São Paulo (with part developed at Yale University). He is a member of the scientific committee of EduLab21 at Ayrton Senna Institute (IAS) He was visiting scholar in 2020 at University of California Berkeley, Institute of Personality and Social Research) CAPES scholarship) and visiting scholar at Lemann Center, Graduate School of Education at Stanford University in 2020. From 2017 he became part of the Advisory Commission on Statistics and Psychometrics of DAEB /INEP (Brazil). He was a member of the questionnaire expert group for PISA 2021 coordinated by the Educational Testing Service (ETS), He is currently member of Technical Advisory Group of OECD study of Social and Emotional Skills, Teaches courses in Psychological Assessment, Psychometrics, Statistical Methods, Item Response Theory and Structural Equation Modeling. Develops research on the Intelligence and Personality Assessment and AI applied to psychometrics.

Wim J. van der Linden

(PRESIDENTIAL ADDRESS)

Wim J. van der Linden is Professor Emeritus of Measurement and Data Analysis, University of Twente, Enschede, The Netherlands.

He is a former Distinguished Scientist and Director of Research and Innovation, Pacific Metrics Corporation, Monterey, CA, and Chief Research Scientist, CTB/McGraw-Hill, Monterey, CA. Dr. van der Linden received his PhD in psychometrics from the University of Amsterdam. His research interests include item response theory, adaptive testing, optimal test assembly, observed-score equating, parameter linking, statistical detection of cheating and response time modeling. He is the author of *Linear Models for Optimal Test Design* (Springer, 2005) and the editor of the three-volume *Handbook of Item Response Theory: Models, Statistical Tools, and Applications* (Chapman & Hall/CRC, 2016, 2018). He is also a co-editor of *Computerized Adaptive Testing: Theory and Applications* (Kluwer, 2000; with C. A. W. Glas), and its sequel *Elements of Adaptive Testing* (Springer, 2010; with C. A. W. Glas). Dr. van der Linden has served on the editorial boards of nearly every major test-theory journal and is co-editor for the Chapman & Hall/CRC Series on Statistics for Social and Behavioral Sciences. He is also a former President of the National Council on Measurement in Education (NCME) and the Psychometric Society, Fellow of the Center for Advanced Study in the Behavioral Sciences, Stanford, CA, was awarded an Honorary Doctorate from Umea University in Sweden in 2008, and is a recipient of the AERA E. F. Lindquist Award as well as the ATP, NCME and Psychometric Career Achievement Awards for his work on educational measurement.

The title of his keynote is *Simplifying large-scale educational assessments through the use of Bayesian adaptive testing*.

S-01: Advancing Adaptive Measurement: Dr. David Weiss's IRT and CAT Lab Innovations in CAT Methodology and Adaptive Measures of Change

This symposium, led by researchers from Dr. David Weiss's IRT and CAT Lab at the University of Minnesota, features four paper presentations detailing novel advancements in adaptive measurement alongside a celebration of Dr. David J. Weiss's foundational influence on the field.

The first presentation explores the complex interactions between CAT scoring methods and stopping rules. Expanding on prior research, this simulation study evaluates four ability estimators alongside four termination criteria across varying item bank configurations. The findings highlight that optimal CAT design is highly contingent on bank quality, demonstrating that while the SEM and fixed-length rules excel in high-information banks, lower-quality banks benefit from a theta change stopping rule paired with weighted maximum likelihood estimation to prevent inefficient testing.

The second presentation shifts to longitudinal applications, focusing on the Adaptive Measure of Change (AMC) for evaluating intra-individual psychometric shifts. Addressing limitations in current pairwise comparison methods, this research introduces a likelihood-based change pattern detection method. By testing constrained alternative models where only a subset of parameters varies, this approach increases the true positive rate and eliminates the need for post-hoc corrections.

The third presentation introduces an equivalence-based framework for evaluating practical differences in Item Response Theory parameter estimation across calibration programs. By simulating datasets based on the graded response model and comparing estimates from legacy and modern software, this research establishes practical thresholds for when observed estimation differences meaningfully impact person-level measurement.

The final presentation grounds these methodological advancements within the broader context of Dr. Weiss's academic legacy. By mapping an academic genealogical tree of Dr. Weiss's research network, this presentation highlights how his extensive collaborations and student mentorship continue to shape the trajectory of modern psychometrics and adaptive measurement methodologies.

Change Pattern Detection in Adaptive Measure of Individual Change

Raj Wahlquist (co-author: David J. Weiss)

Adaptive measure of change (AMC) is a procedure used to test intra-individual psychometric change for some measured latent trait (Finkelman et al., 2010; Tai et al., 2023). By using item response theory based computerized adaptive tests administered at multiple measurement occasions, it is possible to employ null hypothesis significance testing to evaluate significant intra-individual change. This is done using omnibus tests, which have been successfully implemented within the AMC context, and detect when a significant psychometric change has occurred across multiple measurement occasions (Phadke, 2017; Tai et al., 2023). However, these procedures only test if a significant change has occurred between some pair of occasions. When more than two measurement

occasions are present, contrast pairwise comparison methods are required to identify the pair(s) of measurement occasions present in which significant psychometric change has occurred. Recently, methods for performing pairwise comparisons within the AMC context were implemented and examined (Wahlquist & Weiss, 2026). However, these methods relied on using the Benjamini-Hochberg correction (Benjamini & Hochberg, 1995) and there is some concern that the true positive rate (TPR) was disproportionately affected due to this correction. An alternative method is to test the pattern of change, or multiple competing change patterns, directly. Instead of testing the null hypothesis that the estimated trait scores across all measurement occasions are equal against the alternative that all parameters are allowed to vary, it is possible to test a more constrained alternative model in which only a subset of parameters is allowed to vary. Doing so should increase the TPR of the procedure while eliminating the need for post-hoc pairwise comparison testing. When multiple competing models are under consideration, the Akaike Information Criterion and Bayesian Information Criterion (Akaike, 1974; Schwarz, 1978) may be computed for each candidate model based on its likelihood to identify the best supported change pattern. This study applied the likelihood-based change pattern method to AMC and, through a monte-carlo simulation study, compared this method for detecting differences in measurement occasions to those studied by Wahlquist and Weiss (2026).

Interactions Between Termination Criteria and Ability Estimators in Computerized Adaptive Testing

Xinyu Liu (co-author: David J. Weiss)

Computerized adaptive testing (CAT) aims to optimize measurement by tailoring item administration to individual examinees. The efficiency and precision of a CAT heavily depend on the choice of ability (θ) estimator and the termination criterion (stopping rule). Prior research suggests these components interact, but comprehensive evaluations across varying item bank characteristics remain limited. This simulation study investigated the interactive effects of four θ estimators (maximum likelihood [MLE], weighted likelihood [WLE], maximum a posteriori [MAP], and expected a posteriori [EAP]) and four termination criteria (fixed-length, standard error of measurement [SEM], minimum information [MI], and change-in-estimate [$\Delta\theta$]) on measurement bias, precision (RMSE), and test length. These combinations were evaluated across low- (100-item) and high-information (500-item) item banks with both flat and peaked information distributions using the three-parameter logistic model.

The results demonstrated that the optimal CAT configuration is contingent on item bank size and shape. Across all conditions, WLE emerged as the most robust estimator, effectively neutralizing the boundary estimation issues of MLE and the shrinkage bias characteristic of Bayesian estimators. In high-information banks, the SEM and fixed-length rules yielded the lowest conditional RMSE and bias regardless of bank shape. However, in low-information peaked banks, the strict SEM rule frequently failed to reach precision targets at the θ extremes, resulting in inefficient, maximum-length tests. Under these sparse conditions, the $\Delta\theta$ rule paired with WLE provided a superior balance of accuracy and efficiency by halting administration when precision gains stagnated. Conversely, the MI rule consistently exhibited the highest bias and RMSE.

These findings underscore that optimal CAT design is not a one-size-fits-all solution. For high-quality banks, WLE paired with an SEM or fixed-length rule is recommended. For lower-quality banks, practitioners should adopt a $\Delta\theta$ rule or a hybrid SEM approach to prevent inefficient test elongation.

One of these things is not like the other: Setting Thresholds for Equivalence of Item Response Theory Parameter Estimation (Advancing Adaptive Measurement)

Robert Chapman

Psychometric simulation studies often evaluate item response theory calibration software by comparing estimated parameters with known data-generating values using error indices such as root mean square error and bias. Although these indices are useful for assessing statistical recovery, they do not directly indicate whether differences in estimated parameters are large enough to affect applied measurement decisions. This study introduces an equivalence-based framework for evaluating practical differences in graded response model parameter estimates within and across calibration programs.

Using known graded response model parameters for five-category items, 1,000 datasets were simulated with discrimination parameters ranging from 0.5 to 2.5 and boundary parameters ranging from -4 to 4 . Each dataset was calibrated using PARSCALE, Xcalibre 4.2, and the mirt package in R. Estimated item and person parameters were then compared within and across programs to characterize the distribution, magnitude, and ordering of parameter differences. Comparisons included absolute differences in parameter estimates, changes in rank ordering, and associations among estimates using statistics such as Spearman's rho.

The proposed framework shifts the evaluation question from whether a program reproduces the generating parameters to whether alternative estimates are practically equivalent for use in person-level measurement. By summarizing the expected distribution of parameter differences across replications and software programs, this study provides empirical guidance for interpreting when observed differences are negligible, potentially consequential, or large enough to affect calibration, scoring, or adaptive testing applications. This equivalence-based approach offers methodological researchers a complementary criterion for evaluating estimation software and gives applied researchers practical thresholds for judging whether reported software differences are likely to influence scale- or person-level measurement outcomes.

Tracing the Roots and Branching into the Future: The Academic Legacy of Dr. David J. Weiss (Advancing Adaptive Measurement)

Robert Chapman

Dr. David J. Weiss has been a pioneering force in the development of Computerized Adaptive Testing (CAT) and Item Response Theory (IRT). Through decades of mentorship and collaboration, his academic genealogy extends across the field of psychometrics, influencing both theoretical advancements and applied measurement systems. This presentation maps Dr. Weiss's academic family tree, celebrating the widespread impact of his advising and the resulting network of researchers who continue to push the boundaries of adaptive measurement.

Utilizing wide-ranging bibliometric data and graphical analysis of peer-reviewed literature, this work visualizes the scope of his collaborative reach. It systematically traces how his foundational theories on measurement, adaptive models, and intra-individual change have propagated through successive generations of psychometricians. The resulting academic genealogy not only honors a singular career but also illustrates the evolution of CAT methodology as it has been carried forward, adapted, and expanded upon by his students and academic descendants. Ultimately, this presentation highlights the enduring structural impact of dedicated mentorship on the trajectory of modern psychometric research.

S-02: New IRT models to modeling different responses

Item Response Theory (IRT) has evolved substantially beyond its classical formulations, driven by the increasing diversity of response formats encountered in educational, psychological, health, and social science assessments. This symposium presents three recent methodological developments that extend the IRT framework to address important challenges arising from asymmetric item characteristics, ability-based guessing behavior, and count-valued responses.

The first contribution revisits the Skew-Normal IRT family, introducing a Bayesian estimation framework for models capable of accommodating asymmetric item characteristic curves. By relaxing the symmetry assumption underlying traditional logistic and probit models, the proposed approach provides greater flexibility for representing item response processes and offers improved interpretability in applications where asymmetry is present.

The second presentation focuses on a class of ability-based guessing IRT models estimated within a Bayesian framework. Unlike conventional guessing models that assume a constant lower asymptote across examinees, these models explicitly relate guessing behavior to latent ability, allowing for a more realistic representation of response mechanisms. The proposed methodology enhances parameter estimation and provides deeper insights into the interaction between proficiency and guessing strategies.

The third contribution introduces a robust (GNB_k) count IRT model for assessments where responses are naturally recorded as counts. Based on a flexible Negative-Binomial-Gamma distribution, the model accommodates overdispersion, heterogeneity, and outlying observations while preserving an interpretable latent trait structure. Simulation studies and empirical applications demonstrate improved robustness and model fit compared with existing count IRT alternatives.

Together, these three studies illustrate how modern IRT methodology can be adapted to model a wide range of response processes beyond traditional dichotomous data. By incorporating asymmetry, ability-dependent guessing, and count-response mechanisms within unified latent variable frameworks, the symposium highlights current advances that expand the scope, flexibility, and practical applicability of IRT models across diverse assessment contexts.

Revisiting the Skew-Normal IRT family: Bayesian estimation and application

Cristian Bayes (co-author: Jorge Luis Bazán)

The Skew-Normal Item Response Theory (SN-IRT) family, introduced by Bazán et al. (2004), comprises four asymmetric extensions of the traditional symmetric Normal Ogive model. By modeling item characteristic curves with cumulative skew-normal distributions and assigning skew-normal priors to latent abilities, the SN-IRT framework offers enhanced flexibility for dichotomous response data, particularly when ability or item response distributions exhibit asymmetry. This study revisits the SN-IRT family and develops a fully Bayesian estimation framework using modern computational algorithms.

We validate the approach through comprehensive parameter recovery simulations and an empirical application to educational assessment data. Comparative analyses against conventional symmetric and alternative asymmetric IRT models demonstrate the SN-IRT framework's superior model fit, accurate parameter recovery, and robustness to distributional misspecification. These findings highlight the practical and statistical advantages of the SN-IRT family, —offering researchers a flexible, fully Bayesian tool for contemporary psychometric modeling.

A robust GNBk count item response model

Luis Valdivieso

Item response theory (IRT) assumes that responses are influenced by respondents' latent traits and specific item characteristics. IRT models exist now for a variety of different response types, including count data.

Recent Count Item Response Theory (CIRT) models have addressed the limitations of earlier models, thereby broadening their empirical applicability. For instance, Hung (2012) utilized a negative binomial distribution instead of a Rasch's Poisson count model to deal with the overdispersion problem. On the other hand, a flexible Conway-Maxwell-Poisson count model has been proposed by Conway (2002), that handles both overdispersion and underdispersion through the addition of a dispersion parameter. Although, overdispersion is the most common feature found on applications, the possibility to find extreme counting responses and the analysis of how these outliers observations affect estimations, have received little attention on the psychometric literature.

We propose in this work a new and more robust Gamma-Negative-Binomial (GNBk) model, based on Bayes et.al. (2026), that handles not only overdispersion but also deal with the possibility to have extreme responses in the counting process.

Bayesian Estimation for a class of ability-based guessing IRT models

Alex de la Cruz Huayanay (co-authors: Paula Farina, Jorge L. Bazán)

In this study, we present a comprehensive analysis of the Generalized One-Parameter Ability-Based Guessing (G1P-AG) model within a Bayesian framework. This model, which belongs to the family of ability-based guessing models in Item Response Theory (IRT), integrates item characteristics and individual abilities to estimate the probability of a correct response under guessing behavior. Furthermore, it incorporates a general discrimination parameter to accommodate item difficulty.

Our contribution consists of extending this model by adopting a flexible cumulative distribution function (CDF) to define both the item characteristic curve (ICC) for the probability of a correct response and the probability of guessing. Some properties of the proposed model are derived and discussed.

We propose a Bayesian estimation methodology based on the No-U-Turn Sampler (NUTS) and conduct a prior sensitivity analysis to assess the robustness of the inferential

results. The simulation study shows that the proposed algorithm effectively recovers the true parameter values.

Using several model comparison criteria, we apply the proposed methodology to a clinical dataset. The findings highlight the effectiveness of our approach in real-world applications and provide useful insights for psychological and educational assessment practices.

S-03: Frontiers of Bayesian Adaptive Testing

Computerized adaptive testing (CAT) offers substantial psychometric advantages over traditional fixed-form tests, but its practical implementation faces contemporary challenges related to measurement efficiency, large-scale operationalization, and methodological robustness in complex settings. Bayesian Adaptive Testing (BAT) provides a robust framework to address these issues by naturally integrating item parameter uncertainty and supporting advanced psychometric models. This symposium presents four research papers that collectively advance the theory and application of BAT across these critical frontiers.

The first paper by Frey, Schenk, Fink, Liu and König investigates extending BAT to enhance test-taking experience and measurement accuracy simultaneously through the ‘answer-until-correct’ (AUC) response format. This response format allows examinees multiple attempts per item, can mitigate negative test-taking emotions associated with typical CAT implementations without the possibility of response revision and gather more information per item. The study examines the performance of AUC-BAT compared to standard CAT in several scenarios. Initial results confirm that AUC-CAT and AUC-BAT significantly outperform the multiple-choice reference condition, and, crucially, AUC-BAT is shown to reduce bias in ability estimation, especially for examinees with extreme ability levels.

The efficiency of BAT in complex, multidimensional settings is examined in the second paper by Choi and Niu. It compares the spectral efficiency of Fully Bayesian Adaptive Testing (FBAT), which uses joint multidimensional posterior updating and global optimization, with two-level BAT (TLBAT) variants. This research investigates whether efficiency differences stem primarily from posterior factorization or constraints on item-selection geometry under varying levels of latent dimensionality and correlation. The proposed globally optimized TLBAT (gTLBAT) is expected to recover much of the efficiency lost in the sequential TLBAT approach while remaining simpler than FBAT, thus providing a clear information-geometric account of efficiency trade-offs.

The third paper by Patz, van der Linden and Li addresses the operational demands of large-scale assessment systems. A presentation develops an approach for large-scale BAT by focusing on (1) the constraint-governed augmentation of item banks using on-demand item generation from large language models (LLMs), and (2) optimal test assembly using the Shadow-Test Approach within the fully Bayesian framework. This framework, which leverages fast Markov chain Monte Carlo (MCMC) methods, supports real-time updates of student ability and enables unbiased aggregation of results necessary for population-level inference, demonstrating the feasibility of production-ready Bayesian adaptive testing microservices.

The final paper by Templin and Jung addresses a foundational challenge in BAT: the specification of the prior distribution, which is crucial for stabilizing ability estimates, particularly when data is sparse, such as in the early stages of a test. The presentation evaluates five distinct prior specifications, including the standard normal prior, an adaptive normal prior, two novel non-parametric kernel density priors, and a catalytic prior. Findings demonstrate that the optimal prior is assessment-specific, but kernel density priors achieve the best balance of efficiency and accuracy by overcoming the normality assumption and reducing RMSE at the ability extremes, where traditional methods often struggle.

Prior Specification for Ability Estimation in Bayesian Adaptive Testing

Jonathan Templin (co-author: Ae Kyong Jung)

In Bayesian adaptive testing (CAT), the prior distribution is crucial for stabilizing ability estimates, particularly in the early stages where observed data is sparse. Although a standard normal prior (NP) is commonly used in practice, its assumption of normality holds only as the number of items goes to infinity. For any finite test length, particularly the initial items, posterior ability distributions can be skewed, leptokurtic, or platykurtic, rendering the normal distribution implausible for true Bayesian updating. This study addresses these fundamental limitations by empirically evaluating five distinct prior specifications: a static standard normal prior (NP), an adaptive normal prior (NI) updated from the previous stage's posterior mean and variance, two non-parametric kernel density priors (KN, KS), and an application of the catalytic prior (C7).

The kernel density priors (KN and KS) overcome the normality assumption by using kernel density estimation (KDE) to non-parametrically model the prior directly from the examinee's posterior samples, capturing non-normal distributions. The two KDE methods differ by their bandwidth selection: Silverman's Rule-of-Thumb (KN) and the more robust Sheather-Jones plug-in method (KS). The C7 addresses early-stage data sparsity by augmenting the likelihood function. This is achieved by generating a synthetic response for the item providing maximum information at the current ability estimate and incorporating its likelihood into the total likelihood, governed by an adaptive weight that is calculated based on the variance of the previous ability estimate, ensuring the synthetic data's influence diminishes as the estimate becomes more precise (smaller variance).

These priors were compared in a Monte Carlo simulation. The fully crossed factorial design manipulated prior specification, calibration sample size, and test termination rule across 13 true ability levels. Item responses were modeled using the two-parameter logistic (2PL) item response theory (IRT) model. Analysis of the catalytic priors determined that the optimal configuration for maximizing accuracy was the highest catalytic stage combined with the lowest weight tuning parameter.

Results revealed a critical and consistent trade-off between measurement precision and test efficiency across the five priors. The priors clustered into three distinct performance profiles. The NP and C7 priors maximized precision, yielding the lowest and most stable standard errors across the entire ability spectrum, but required the longest tests. The C7 prior performed nearly identically to the NP prior in accuracy but demonstrated a marginal efficiency gain, making it an option for high-precision assessments. In contrast, the NI prior was a pure efficiency model, resulting in the shortest tests, but produced the highest overall RMSE and bias. This was characterized by an inverted U-shaped SE curve, suggesting the prior became overly confident in inaccurate estimates at the ability extremes.

Occupying the middle ground, the kernel density priors (KN, KS) were efficient, requiring fewer items than NP/C7. They offered the best balance, combining high efficiency with the lowest RMSE at the ability extremes, demonstrating better accuracy where traditional methods struggled. The findings demonstrate that no single prior is universally optimal; the best choice is contingent on aligning the prior function with the specific goals of the assessment.

Spectral Efficiency in Multidimensional Bayesian Adaptive Testing: Comparing Fully Bayesian, Sequential Two-Level, and Globally Optimized Two-Level Frameworks

Seung Choi (co-author: Luping Niu)

Bayesian methods in computerized adaptive testing are increasingly being extended to multidimensional settings, where relationships among latent traits can be leveraged to improve measurement efficiency. Two prominent approaches are fully Bayesian adaptive testing (FBAT), which performs joint multidimensional posterior updating with globally optimized item selection, and two-level Bayesian adaptive testing (TLBAT), a computationally simpler framework composed of multiple conditionally unidimensional adaptive tests linked by latent trait correlations. In its classical form, sequential TLBAT (sTLBAT) administers dimensions one at a time and rescales previously estimated dimensions only after all dimensions are completed. Although FBAT and sTLBAT may rely on the same latent covariance structure, they differ fundamentally in how information propagates across the eigenspace of latent trait uncertainty.

This study develops a unified theoretical and simulation framework to compare FBAT with both sTLBAT and a proposed globally optimized TLBAT variant (gTLBAT). To ensure a principled comparison of adaptive information allocation, all frameworks use mutual information (MI) for item selection. Unlike sTLBAT, gTLBAT relaxes the sequential administration constraint, allowing items from any dimension to be selected at each step based on their expected reduction in global posterior entropy, while retaining conditionally unidimensional posterior updating. The study investigates whether differences in adaptive efficiency arise primarily from posterior factorization, constrained item-selection geometry, or their interaction.

Item responses are generated using between-item multidimensional item response theory models with simple structure and correlated latent traits, under varying dimensionality ($K=3, 5, 7$) and magnitudes of latent correlations. Two complementary evaluation regimes are considered. In an oracle covariance design, the latent covariance matrix (Σ) is fixed at its true generating value across all methods, isolating the intrinsic algorithmic efficiency of each adaptive framework under known latent geometry. In a full-system design, each framework estimates the latent covariance structure using its own inferential assumptions, allowing evaluation of realistic psychometric performance under covariance uncertainty.

Performance is evaluated using latent trait recovery, posterior entropy reduction, information gain per administered item, and recovery of the eigenspace structure of Σ . The study is motivated by a spectral theory of adaptive testing efficiency, which predicts that FBAT will increasingly outperform sTLBAT as latent trait correlations strengthen and the dominant eigenvalue of Σ becomes more pronounced. The proposed gTLBAT framework is expected to recover a substantial portion of the efficiency lost by sTLBAT through globally optimized item selection, while remaining asymptotically less efficient than FBAT because posterior updating remains factorized across dimensions. Collectively, the study provides an information-geometric account of efficiency in multidimensional Bayesian adaptive testing and clarifies the conditions under which

computationally simpler two-level frameworks can approximate the performance of fully Bayesian methods.

Bayesian Adaptive Testing for the Answer-Until-Correct-Response Format

Andreas Frey (co-authors: Cosima Schenk, Aron Fink, Tuo Liu, Christoph König)

Computerized adaptive testing (CAT) is advantageous from a psychometric point of view. At the same time, meta-analytic results suggest that CAT can increase negative test-taking emotions when revising given responses is not allowed, which is the case in almost all CAT implementations. One way to make item review in CAT possible, is to use the answer-until-correct (AUC) response format (AUC-CAT). AUC allows examinees to attempt an item multiple times until a correct response had been given. In addition to the potentially more favorable emotional test experience, AUC-CAT is also psychometrically advantageous. Since more than one response can be collected per item, AUC-CAT can gather more information compared to the multiple-choice response format, which is likely to increase the precision of ability estimation. The individual attempts taken in AUC can be parameterized with sequential IRT models (SIRT). As a result, more item parameters are estimated when using SIRT models compared to dichotomous IRT models such as the 1-2-, or 3PL. Thus, it is very important to account for the statistical uncertainty of the item parameter estimates in adaptive item selection and ability estimation. Otherwise, capitalization on item calibration error is to be expected, leading to biased ability estimation particularly for extreme ability levels. The statistical uncertainty of the item parameter estimates can be taken into account through fully Bayesian adaptive testing (BAT).

This study is the first to examine the performance of BAT with AUC (AUC-BAT) with a comprehensive simulation study. The study uses a 2x2x2x2 factorial design with the factors response format (AUC, MC), Bayesian (BAT, CAT), responses per item during item pool calibration (500, 1,000), and item pool difficulty (empirical/hard, average). Average Bias and RMSE of the ability estimates and observed test information were calculated as dependent variables for 11 ability steps from -2.5 to 2.5. For item selection, a generalization of the approach of van der Linden and Ren (2020) was adopted for BAT, and the maximum information criterion for CAT. The item pools contained 100 items with four response categories from which 20 items were adaptively presented to each simulee. The simulation was conducted using the R packages mirt, mirtCAT, and CmdStanR interfacing Stan 2.37. The analyses are currently running but will well be completed by the time of the conference.

First results show that (1) AUC-CAT and AUC-BAT clearly outperform the reference condition MC-CAT, (2) Bias is lower for AUC-BAT than for AUC-CAT especially for extremes ability levels, and (3) AUC-CAT achieves somewhat lower RMSE and somewhat higher observed test information than AUC-BAT. The results show that using AUC in adaptive testing is highly efficient and that adopting a fully Bayesian approach (AUC-BAT) is capable to reduce bias of ability estimation with only a slight decrease in measurement precision.

Bayesian Optimal Large-Scale CAT Using Constraint-Governed Augmentation

Richard Patz (co-authors: Wim J. van der Linden, Jie Li)

Educational assessment increasingly occurs within interconnected learning and credentialing ecosystems, yet high-stakes summative testing programs remain operationally and psychometrically disconnected from instructional, practice, and formative assessment environments. In both K–12 accountability systems and professional licensure contexts, the economic and security requirements of validated item pools create tensions between content integrity, transparency, instructional coherence, and scalability.

Efforts to create comprehensive assessment systems that bring coherence across learning, practice, and summative experiences have advanced in recent years. A conceptually clear design objective in these solutions is a large bank of high-quality test items, tagged to detailed learning, performance, or competency frameworks, and a content delivery algorithm that ensures educational objectives are met under conditions that assure instructional coherence and psychometric integrity. Economic constraints make such ideal solutions generally infeasible under most circumstances, requiring careful design to balance competing objectives under practical resource limitations.

In this paper we develop an approach to adaptive testing and item bank management that meets the requirements of comprehensive large-scale assessment systems by focusing on two technical challenges: (1) the constraint-governed, algorithmic augmentation of item banks with on-demand item generation leveraging large language models (LLMs), and (2) the assembly of forms (including adaptive forms) and production of ability estimates that satisfy measurement objectives and constraints at both individual and aggregate targets of inference.

Recent advances in fast Markov chain Monte Carlo (MCMC) methods for Bayesian CAT now make it possible to update posterior distributions of student ability in real time, yielding immediate draws from individual posterior distributions that support both adaptive item selection and unbiased aggregation of results at higher levels of inference. How this framework can support the optimal assembly of collections of forms for population-level inference is discussed.

We employ optimal test assembly for adaptive testing (i.e., the Shadow-Test Approach) within a fully Bayesian CAT framework, demonstrating the feasibility of the approach using a production-ready Bayesian adaptive testing microservice implementing shadow-test optimization.

S-04: Continuous Item Banking for a High-Stakes Certification Without Pretesting: An Operational Decomposition

In February 2026, the Brazilian Financial and Capital Markets Association (ANBIMA) implemented a new architecture for its investment distribution certifications, replacing fixed exam booklets with per-candidate-unique forms drawn dynamically from a shared item bank. The change exposes a tension absent from the textbook computerized adaptive testing literature: in a high-security financial certification context, items cannot be pretested in any controlled environment prior to operational use. Every item enters the bank without prior calibration and must be evaluated inline as candidates take the exam. This forces a re-engineering of three intertwined operational components — how new item formats are designed, how the bank is monitored, and how items are commissioned to replenish it — that must work jointly rather than in isolation.

This symposium presents the operational system deployed for ANBIMA's CPA certification, with validation from 8,115 candidates and 464,800 item responses observed between 19 February and 14 May 2026. The three presentations follow the operational sequence: item design, quality monitoring, demand-driven replenishment.

The first presentation addresses how to deliver branching-scenario items — ANBIMA's Questões Interativas format — in a high-security exam where exam-security constraints exclude generative AI from delivery. The Questões Interativas bundle (a five-decision árvore de diálogo com etapas pré-fixadas plus two contextual MC items, all scored dichotomously) is validated against the two standard reviewer objections: bundle discrimination (corrected point-biserial 0.41 vs 0.26 for ME, $p < 0.001$) and within-bundle local independence (Yen's Q3 mean 0.019, empirically indistinguishable from a null of random pairs).

The second presentation introduces a six-dimensional flagging framework over Classical Test Theory metrics for monitoring items entering the bank without pretesting. Bank reliability climbed monotonically from $\alpha = 0.43$ in operational week 1, crossed the certification-grade $\alpha \geq 0.90$ threshold by week 4, and stabilized at $\alpha \geq 0.95$ from operational week 9 onward as the calibrated item count saturated at 305 of 333 bank items (91.6%). A Joint-MLE Rasch (1PL) cross-check on the 293 calibrated multiple-choice items yields a Spearman rank correlation between Rasch difficulty and CTT proportion-correct of -0.9987 , defusing the recurrent IRT-vs-CTT concern in non-pretested-bank contexts.

The third presentation introduces a formal item-demand function: the continuous, operational counterpart to van der Linden and Boekkooi-Timminga's (1989, 1992) static integer-programming bank design. Demand is ranked across (content node $\times$ format $\times$ difficulty band) sockets by a decomposed signal combining exposure-weighted inverse-quality with a symmetric band-pressure multiplier clamp (target/current, 0.1, 5.0) that amplifies under-represented bands and dampens over-represented ones. Under simulation over 100 demanded items, the formulation closes 21% of the L1 gap to target band distribution, outperforming exposure-only, quality-only, and uncoordinated baselines.

The integrated argument: an item-banking pipeline that cannot rely on pretesting can nevertheless achieve and sustain certification-grade reliability if its item-design,

monitoring, and demand components are designed as a closed loop, consistent with the continuous-processes paradigm for modern adaptive testing articulated by van der Linden (IACAT 2022 keynote). The CPA case provides empirical evidence for each component, and methodological generalizations are discussed for high-stakes certification programs operating under analogous security constraints.

Interactive Branching-Scenario Items in High-Stakes Financial Certification: Bundle Scoring Without Local Dependence

Andrea Burgos de Azevedo Mangabeira (co-authors: Daniel Munduruca, Marco Pepe)

ANBIMA's CPA professional certification — mandatory for distributing financial products to retail clients in Brazil — must measure cognitive demands that mirror real client-advisor interactions, while operating under exam-security constraints that exclude generative AI from delivery. This presentation describes the design, scoring, and psychometric validation of ANBIMA's Interactive Items (II), item format, deployed at scale in CPA from February 2026 under the New Assessment Architecture. Technically, each II is a dialogue tree with pre-determined phases — a deterministic dialogue tree authored in advance, with no AI-generated content at delivery time.

Each Interactive Item bundle is a 3-item testlet: a five-decision branching dialogue scenario scored dichotomously on whether the candidate reaches the correct terminal node, plus two multiple-choice items that use the same dialogue context. The choice to score the dialogue dichotomously rather than as a polytomous response pattern is operationally motivated by the certification's absolute 70%-correct pass criterion and the production constraint of fully rubric-based scoring without an IRT model. The design hypothesis is that scenario fidelity can be preserved without invoking generative AI by structuring the interactivity into a fixed branching graph with pre-fixed stages.

We evaluate the design on 8,115 CPA candidates (19 February – 14 May 2026) who saw at least one of the 12 currently calibrated QI bundles.

Discrimination. Four point-biserial variants are computed at both sub-item and bundle granularity, with each variant addressing the proportion imbalance between QI and multiple-choice (ME) items in any given booklet. At sub-item level, corrected QI discrimination is statistically equivalent to ME (BIS 0.23 vs 0.26, $p = 0.19$). At bundle level (the 3 sub-items aggregated into one score per candidate), corrected QI discrimination substantially exceeds ME (0.41 vs 0.26, $p < 0.001$), consistent with the Spearman-Brown gain expected if local independence is approximately satisfied within the bundle.

Local independence. Yen's Q3 residual correlations from a Rasch (1PL) fit on the sub-item response matrix yield within-bundle mean $Q3 = 0.019$ (std 0.062), with 35 of 36 within-bundle pairs below the conventional $|0.20|$ threshold. A null distribution constructed from 733 random ME-ME, QI-QI-cross-bundle, and QI-ME pairs has mean $Q3 \approx -0.02$ with higher dispersion (std 0.08-0.11). A one-tailed Mann-Whitney test of (within $>$ null) yields $p = 0.997$: within-bundle $Q3$ is empirically indistinguishable from random pairs, with a tighter distribution. The bundle design does not induce local dependence above the noise floor.

Cognitive demand. Interactive items are concentrated at higher Bloom-revised levels by design: 33% of bundles are at NC4 (Analyze) compared to 22% of calibrated ME items, with the remaining bundles at NC2.

We argue that interactive branching-scenario items with pre-fixed stages can be deployed at scale in high-stakes certification without invoking generative AI or polytomous testlet IRT, provided the bundle is treated as a unit and local independence is empirically verified rather than assumed.

Closing the Loop on Item Demand: A Decomposable Formalism for Continuous Item Banking

Marco Pepe (co-authors: Andrea Burgos de Azevedo Mangabeira, Wesley Tenório)

Item bank design has been formalized as a static integer-programming problem since van der Linden and Boekkooi-Timminga (1989, 1992): minimize the cost of writing items subject to test information function targets and blueprint constraints. The operational counterpart — what to commission next once a bank is in production, drifting out of calibration and exposure equilibrium — has remained largely informal. This presentation introduces a formal item-demand function deployed at ANBIMA's CPA certification under its New Assessment Architecture, validated by simulation against principled counterfactual strategies. The contribution is a continuous, operational counterpart to van der Linden and Boekkooi-Timminga's static design-phase formalism, integrating quality monitoring (cf. Gianopoulos et al., 2025) with distributional steering rather than treating cost minimization alone.

The formalism. For each socket $s = (\text{Topic} \times \text{Format} \times \text{Difficulty band})$, demand is ranked by $\text{Estresse}(s) = \text{quality_stress}(s) \times \text{band_pressure}(\text{band})$. The first factor, $\text{quality_stress}(s) = \sum_i \text{EXP_ADJ}_i \times (1 - f(\text{BIS}_i)) / \sum_i (1 - f(\text{BIS}_i))$, is exposure-weighted inverse-quality: items with high exposure and low discrimination contribute more pressure on the socket they occupy. The second factor, $\text{band_pressure}(\text{band}) = \text{clamp}(\text{target}(\text{band})/\text{current}(\text{band}), 0.1, 5.0)$, is a symmetric multiplier on band representation: under-represented bands are amplified up to 5×; over-represented bands are dampened down to 0.1×. The product directs demand to sockets where both quality erosion and distributional gap are concurrent.

Empirical validation. An iterative simulator adds $N = 100$ hypothetical demanded items to the highest-stress socket sequentially, recomputing the stress signal after each addition, and measures L1 distance between the bank's resulting band distribution and the production target (CPA: 60% Easy, 35% Medium, 5% Difficult). Starting from the observed state on 15 May 2026 (45% / 37% / 18%, $L1 = 0.30$), 100 simulated demands following stress drop L1 monotonically to 0.24 — a 21% gap closure, with the band distribution converging toward target.

Counterfactual comparison. Three principled alternatives are evaluated. Pure-inverse-quality demand concentrates on the difficult band, where BIS is systematically lower due to range restriction in the pre-selected candidate population, doubling difficult share to 36% and driving L1 to 0.62 — the opposite of the intended steering. Uncoordinated demand, sampling bands from the current distribution, leaves L1 essentially unchanged at 0.30. Pure exposure-based ranking incidentally lowers L1 to 0.17 on this snapshot because the highest-exposure sockets happen to fall in the most under-represented band,

but pure exposure has no representational mechanism to stop demanding new items once a bank exceeds its target and would overshoot under longer horizons.

Implications. Decomposing the demand signal into quality monitoring and distributional steering — multiplicatively, with each factor mathematically symmetric in its handling of over- and under-representation — makes explicit a design choice operational programs frequently conflate. The formulation demonstrates measurable convergence toward target distribution while preserving the quality signal that drives item revision and suspension decisions downstream. The approach is generalizable to other high-stakes certification programs that must replenish a bank under live exam load without prior pretesting.

Continuous Calibration of an Operational Item Bank Without Pretesting: A 13-Week Field Test

Erica Ruiz (co-authors: Marco Pepe, Andrea Burgos de Azevedo Mangabeira, Wesley Tenório)

ANBIMA's New Assessment Architecture, operational since February 2026, draws each candidate's exam booklet uniquely from a shared item bank. Combined with security constraints that prohibit pretesting items in any controlled environment, this configuration requires the bank to be calibrated continuously as candidates take the live exam. This presentation describes the monitoring framework deployed to manage this configuration and reports empirical results from the first 13 weeks of operation for the CPA certification.

The framework. Items are evaluated inline using Classical Test Theory metrics on the sparse per-candidate response matrix. Six quality flags are computed continuously per item once item exposure crosses $N = 250$ distinct respondents: extreme proportion correct ($p \geq 0.95$ or ≤ 0.25), low discrimination (item-removed point-biserial BIS < 0.10), negative discrimination (BIS < 0), non-functional distractors (option chosen by $< 5\%$ of candidates), high omission rate ($> 5\%$), and content-leakage suspicion ($\Delta p > 5$ percentage points between first and last rolling 360-candidate windows). Bank-level reliability is estimated using Cronbach's α with pairwise covariance and minimum overlap of 30 — the appropriate treatment for the sparse matrix produced by unique-per-candidate booklets, where each candidate sees only a subset of items, making the standard complete-data α formula inapplicable.

Empirical trajectory. On 8,115 CPA candidates (19 February – 14 May 2026, 464,800 item responses), bank reliability climbed monotonically from $\alpha = 0.43$ in week 1 (227 candidates, 5 calibrated items), through $\alpha = 0.79$ in week 2, $\alpha = 0.91$ in week 4 (crossing the certification-grade $\alpha \geq 0.90$ threshold), $\alpha = 0.94$ in week 6, and stabilized at $\alpha \geq 0.95$ from week 9 onward as the calibrated item count saturated at 305 of 333 bank items (91.6%). Bank reliability crossed the certification-grade threshold within four operational weeks without any prior calibration phase, and the $\alpha \geq 0.95$ plateau has held through week 13.

IRT cross-check. A Joint Maximum Likelihood Rasch (1PL) fit on the 293 calibrated multiple-choice items (mean infit MNSQ = 0.996, mean outfit = 0.976, 99.3% of items within the [0.5, 1.5] fit window) yields a Spearman rank correlation between Rasch difficulty b and CTT proportion-correct of -0.9987 (Pearson -0.976). The CTT-based monitoring framework recovers the same item-difficulty ordering that an IRT calibration would, with substantially lower data requirements per item. The correlation between

Rasch b and the corrected point-biserial discrimination is weak (-0.24), as expected when comparing a difficulty parameter against a discrimination metric; the modest negative direction is consistent with mild range restriction in the pre-selected candidate population. Implications. A non-prettested bank managed by continuous CTT-based monitoring reaches certification-grade reliability within four to five operational weeks and stabilizes thereafter. The framework surfaces actionable item-level alerts (46 calibrated items currently flagged for non-distractor reasons, including 6 with negative discrimination scheduled for immediate suspension) that feed directly into the content team's review workflow. The approach is potentially generalizable to other high-stakes certification programs where pretesting is prohibited or impractical.

S-05: The PROMIS, NIH Toolbox, and Mobile Toolbox Measurement Systems: Architecture, Psychometric Foundation, and the Challenges of Adaptive Health Measurement

This symposium introduces three major health measurement systems used in medical assessment across over 1,100 institutions globally: PROMIS, NIH Toolbox for the Assessment of Neurological and Behavioral Function (NIHTB), and Mobile Toolbox (MTB) for Monitoring Cognitive Function by outlining both the structural architecture and the psychometric principles underlying health measurement, and introduces a series of operational research questions and empirical solutions for each measurement system, highlighting its distinctive measurement contexts.

The Patient-Reported Outcomes Measurement Information System® (PROMIS®) was established in 2004 with funding from the National Institutes of Health (NIH) as part of the NIH Roadmap for Medical Research. PROMIS comprises a system of person-centered measures designed to evaluate and monitor physical, mental, and social health in both adults and children. It was developed to address a fundamental limitation of traditional patient-reported outcome (PRO) instruments: their inability, under classical test theory, to achieve precision, brevity, and comparability simultaneously across diverse populations and chronic conditions. The NIHTB is a large battery of instruments including both PRO and performance measurement across four domains: cognitive, motor, sensory and emotional health, and incorporates millisecond-level timing in the scoring algorithms. MTB is an expandable, open-source library of smartphone-based tests to measure cognitive, emotion, and increasingly

All of the health measurement systems rely on large item banks for each health and cognitive domain, such as pain interference, fatigue, physical functioning, emotional distress, and social role participation, calibrated using item response theory (IRT). This approach allows measures to be administered flexibly while remaining directly comparable across different short forms and CAT administrations. Samejima's graded response model (GRM) serves as the primary IRT framework, reflecting the polytomous, Likert-type format of PRO items. For example, all PROMIS scores are reported as T-scores standardized to the U.S. general population (mean = 50, standard deviation = 10), enabling cross-domain and cross-disease comparisons that are not typically possible with legacy PRO measures.

Measures are available in multiple formats, including short forms (fixed sets of 4–10 items) and CATs that dynamically select items based on prior responses, and multi-domain profiles. The CAT engine operates through a sequential adaptive process: an initial θ estimate is generated, the most informative item is selected to refine that estimate, the score is updated after each response, and administration continues until predefined stopping criteria are met. Several stopping rules are available for different clinical contexts, including for PROMIS, a “screen-to-CAT” option with an early stopping condition: if a respondent selects the best health response on the first item, the assessment terminates, a method primarily intended for screening rather than precise estimation.

Finally, we situate PROMIS, NIH Toolbox, and MTB within the broader landscape of health measurement (PRO and performance measurement) by highlighting key distinctions from educational testing, including the nature of the latent trait (self-reported experiential health status), the importance of face validity and content representation in item selection, and the implications of floor and ceiling effects in clinically extreme populations. We close implications for longitudinal, clinical and comparative effectiveness research.

Computerized Adaptive Testing (CAT) for Patient-Reported Outcomes Measurement Information System (PROMIS): Evaluation of Stopping Rules for the Pediatric Item Banks

Jiwon Kim (co-authors: Xiaodan Tang, Jin-Shei Lai, Benjamin Schalet, Emily Ho, Richard Gershon, David Cella)

The Patient-Reported Outcomes Measurement System (PROMIS®) provides a wide range of person-centered measures for physical, mental, and social health in adults and children, designed for use across both general and clinical populations. Within the PROMIS item banks, computerized adaptive testing (CAT) reduces respondent burden by tailoring item administration to each respondent while maintaining measurement precision. While decisions on the stopping rules of CAT administration can meaningfully affect scores, test length, and reliability, the consensus on the optimal stopping rule across the pediatric health domains remains limited.

This study evaluated different stopping rules for the PROMIS Pediatric items banks for fatigue, depression, anxiety, anger, peer relationships, mobility, and upper extremity function. CAT simulations were conducted in R, with 1,000 simulees generated per item bank from a normal distribution (mean = 0, SD = 1). The simulation conditions for the stopping rules included: maximum number of items administered (8 or 12), minimum number of items administered (4 or 5), standard error (SE) criteria (≤ 0.3 or ≤ 0.4), and SE reduction (SER) from the previous iteration (none or ≤ 0.01).

A Screen-to-CAT approach was also examined, where administration stops after the first two items if the healthiest possible responses are endorsed. Evaluation outcomes included average test length, T-score range, mean SE, and the proportion of scores achieving reliability greater than 0.80. Across all item banks, the stopping rule of minimum 4-item, maximum 8-item, and SE ≤ 0.30 produced score ranges, mean SEs, and reliability levels comparable to those obtained using a lengthier rule (minimum 5-item, maximum 12-item, and SE ≤ 0.40), while requiring fewer items on average. Adding a SER ≤ 0.01 criterion reduced the average test length, but at the same time narrowed the score ranges, increased the mean SEs, and decreased the proportion of scores with reliability greater than 0.80. However, the differences were minimal. Given that patient-reported outcome measures are not diagnostic and therefore are generally regarded as lower-stakes than educational assessments, we recommend a minimum 4- and maximum 8-item default setting. For Screen-to-CAT, selecting the healthiest item as the first screening item produced higher proportions of reliable scores than selecting the average item across the

simulation conditions. These findings support a stopping rule of minimum of 4-item and maximum 8-item CAT assessment, and either SE ≤ 0.30 or SER ≤ 0.001 (whichever occurs first) for CAT administration of the PROMIS pediatric item banks. The same rules, together with the best health starting item, are recommended for Screen-to-CAT administration. Establishing standardized, evidence-based stopping rules across the pediatric PROMIS item banks offers practical guidelines for more efficient and consistent measurement in both research and clinical settings.

NIH Toolbox Calibration and Scaling for a multi-stage episodic memory test across the lifespan

Emily Ho (co-authors: Erica LaForte, Aaron Kaat, Julie Hook, Catherine Han, Richard Gershon)

The NIHTB Picture Sequence Memory test, a measure of episodic memory, consists of learning, then recalling a series of pictures related to one of two themes (either Fair or Park), with a pre-specified number of sequences (3, 6, 9, or 15 pictures). To develop a multi-stage test (MST) for assessing episodic memory across the lifespan, we conducted a large-scale norming study of over 3900 nationally representative individuals ages 3 and up to evaluate and determine 1) the empirical difficulty of each sequence length; and 2) the equivalence of the two Park and Fair theme sequences. We describe the norming design for the routing sequence (Table 1), where norming participants were randomly divided into one of two Forms containing pre-specified routing sequences to ensure adequate person data per theme and sequence. Using a randomly equivalent groups design, all item data from each examinee were used for item calibration. We describe the process to estimate, using a Rasch partial-credit model, all person ability estimates from the two forms so that they are on the same scale. Figure 1 shows the resulting relative difficulty of each of the sequences after the co-calibration. There are equivalent raw-score difficulties for the two themes at the longer sequence lengths (i.e., 9 and 15), and more differential difficulty as sequences become briefer, likely due to a floor effect for younger children being administered the shortest sequences. We discuss how the calibration process informed the final workflow of the MST to ensure the youngest children can be assessed whilst mirroring the MST routing scheme deployed in the norming study.

Computer Adaptive Testing on Mobile Devices: Assessments of Verbal Ability for Remote Administration in Mobile Toolbox

Y. Catherine Han (co-authors: Elizabeth M. Dworak, Aaron J. Kaat, Jiwon Kim, Augusta Zukauskas, Emily Ho, Michael Weiner, Joseph Eichenbaum, Bernard Landavazo, Monica R. Camacho, Rachel L. Nosheny, Richard C. Gershon)

As modern healthcare and research become increasingly digitized, routine remote monitoring of cognitive function is critical for improving accessibility of standardized assessments. The Mobile Toolbox (MTB), a smartphone-based suite of tests publicly released in 2024 and freely available via REDCap's MyCap App, represents recent

advancements in remote health and cognitive monitoring. The MTB cognitive tests measure well-established neuropsychological constructs, such as receptive language, executive function, memory, and processing speed. Within the MTB, two Computer Adaptive Tests (CATs) measuring verbal ability—Word Meaning and Spelling—are calibrated using Item Response Theory. While the validity and test-retest reliability of these measures have previously been demonstrated in adults aged 20-85 (Young et al., 2024; LaForte et al., 2024), longitudinal studies require a large, diverse item pool and item exposure management to maintain high-precision measurement over repeated administrations.

A sample of $n=576$ English-speaking adult participants from the Brain Health Registry study (baseline: age $M=62.6$ years, $SD=12.6$, 81.9% female) was assigned to complete the full MTB battery remotely on their personal iOS smartphones, including Word Meaning and Spelling. A modified Simpson-Hetter exposure control method targeted a 60% administration rate per item. Examinees were required to complete a minimum number of live (non-practice) items (Spelling=10; Word Meaning=15). After the minimum was reached, testing terminated when a prespecified standard error (SE) threshold was achieved (approximating a person-specific reliability of 0.75) or when the maximum number of items was administered (30 items for both).

At baseline, median completion times were 4.43 minutes ($IQR=1.92$) for Word Meaning and 4.52 minutes ($IQR=1.78$) for Spelling. At retest, median times decreased to 3.94 minutes ($SD=1.75$) for Word Meaning and 3.74 minutes ($IQR=1.78$) for Spelling. Retest times were significantly shorter for Word Meaning and Spelling (paired Wilcoxon signed-rank tests, $p<.001$), suggesting that using prior ability estimates at retest improved administrative efficiency and reduced participant burden. Median test lengths were nearly identical across sessions (Word Meaning=23 items, $IQR=3$; Spelling=10 items, $IQR=0$).

Most participants terminated the session after reaching the SE criterion (Word Meaning: 72.3% baseline, 71.0% retest; Spelling: 98.6% baseline, 98.7% retest), with the remainder reaching the maximum number of items. Mean final conditional SEs at baseline were 0.51 ($SD=0.03$) and 0.38 ($SD=0.04$) for Word Meaning and Spelling, respectively. At retest, mean SEs were 0.51 ($SD=0.03$) for Word Meaning and 0.39 ($SD=0.03$) for Spelling. Item exposure rates closely matched the 60% target. At baseline, only 0.41% (Word Meaning) 0.69% (Spelling) of the items exceeded the target rate. At retest, 0.25% (Word Meaning) and 0.65% (Spelling) exceeded the target, indicating that item administration under repeated remote deployment was effectively regulated.

Overall, CAT performance demonstrated stable precision, tightly controlled item exposure, and improved retest efficiency. These results support the feasibility of scalable, longitudinal remote cognitive monitoring using smartphone-based adaptive testing. Future work will investigate if these findings differ as a function of retest interval, including potential efficiency gains, precision, and exposure patterns across intervals. Cross-session item overlap will also be evaluated with respect to score changes to more directly assess practice effects.

S-06: EduCAT

Adaptive Assessment in Practice: Applications of CAT at SESI and in the ANPAD Test

Franciele Sena

The lecture will present concrete experiences of applying Computerized Adaptive Testing (CAT) in Brazil, highlighting its use in SESI assessments in Elementary and High School and in the ANPAD Test. The functioning of the methodology, the personalized selection of items according to each participant's performance, and the benefits achieved, such as greater precision, agility, and efficiency in assessment, will be discussed. The presentation will also address the lessons learned and challenges involved in implementing CAT in different educational contexts.

Item Calibration with Avatars: Innovation and Equity in Assessment

Dalton F. De Andrade

The lecture will explore the use of Avatars in the item calibration process, showing how simulated participants can support the development and improvement of assessment instruments. The statistical fundamentals of calibration, the possibilities of using Avatars to simulate different proficiency levels and response profiles, and the care needed to ensure validity, precision, and equity will be presented. The proposal highlights how Statistics, Technology, and Artificial Intelligence can contribute to more robust, efficient, and innovative assessments.

Individual Presentations

Development and Evaluation of an IRT-Based Computerized Adaptive Testing System for General English Proficiency of College Students

Dong Seo (co-author: Hyun Park)

The purpose of this study was to develop and evaluate a computerized adaptive testing (CAT) system for efficiently and accurately diagnosing the general English proficiency of college students. To construct the CAT system, an item bank consisting of 100 TOEIC practice items (50 listening and 50 reading items) was calibrated using item response theory (IRT). Based on the calibrated item parameters, the CAT system was designed to support both fixed-length and variable-length testing formats. To examine the effectiveness of the developed system, 60 college students enrolled in two sections of a general English course at Hallym University participated in the study. The effectiveness of the CAT system was evaluated by comparing learning outcomes between the CAT group and the non-CAT group. The results showed that, in the listening domain, the CAT group demonstrated significantly greater score improvement than the non-CAT group, indicating a significant interaction effect. In the reading domain, both groups showed score gains over time; However, statistically significant improvement was observed only in the CAT groups based on simple effects analysis. These findings suggest that the IRT-based CAT system can provide both measurement precision and administrative efficiency in diagnosing general English proficiency. Furthermore, the results imply that repeated use of the CAT system as a supplementary learning tool may positively contribute to the improvement of English language ability.

Growth Norms for Computer Adaptive Tests: Evaluating Three Conditional Growth Percentile Methods

Luciana Cancado (co-authors: Logan Rome, Jiawei Xiong)

Normative growth reporting, where a student's growth is compared to that of students with similar starting scores, is a high-demand feature of operational computer adaptive tests (CAT) used in educational settings. This study compares three conditional growth percentile methods: Student Growth Percentiles (SGP; Betebenner, 2009), Percentile Rank of Residuals (PRR; Castellano & Ho, 2013, 2015), and the less explored and not yet formally named Theoretical Gain Percentiles (TGP; derived from Wyse & Seo, 2014), applied to an operational CAT-based interim assessment administered to millions of students in the United States (U.S.).

Methods

Data. Fall–spring math interim assessment data from 2024–25 (grades K–8), constructed to be nationally representative of the U.S. population across region, urbanicity, ethnicity, and income.

Method. For PRR and TGP, ordinary least squares (OLS) regression models predict spring scores from fall scores by grade. Expected Growth Values (EGVs) are derived as the difference between theoretical starting scores based on the full assessment scale and model-predicted ending score.

TGP assumes normally distributed residuals (mean = 0, SD = model RMSE) and uses the normal cumulative density function to derive the gain score at each (1-99) percentile rank. PRR derives percentile ranks empirically from observed residuals, then maps them to gain scores via EGVs.

SGP uses quantile regression as implemented by the R SGP package to estimate conditional growth percentiles.

Similar to Wyse & Seo (2014), percentile ranks produced by each method were pairwise compared using four criteria: (1) correlations (2) root mean square difference (RMSD), (3) mean difference (MD), and (4) classification consistency determined by the percent agreement for the two methods in terms of classifying students as exhibiting low (< 35th percentile), typical (between 35th and 65th percentiles), or high (> 65th percentile) growth.

Results

Residual plots (Figure 1) indicate mild heteroscedasticity, challenging the homoscedasticity assumption underlying PRR and TGP. Figure 2 shows a comparison of gain score distributions for the 50th percentile of growth across methods by starting score. We see that TGP and PRR values are closely aligned, while SGPs diverge at points. Table 1 shows pairwise comparisons of the three different conditional growth percentile methods across the four criteria. Correlations ranged from 0.98 to 0.99. TGP and PRR showed consistently lower RMSD values relative to each other than either did to SGP, reflecting their shared regression foundation. PRR percentile ranks were systematically higher than SGP on average. Classification consistency ranged from 87.3% to 95.8% across all grades and method pairs.

TGP–PRR agreement was strongest overall, as expected given the shared model. SGP, despite its nonparametric approach, remained highly correlated and consistent with both OLS-based methods. Results suggest practitioners can treat these methods as largely interchangeable in operational contexts, even under mild violations of regression

assumptions—a finding relevant to adaptive testing systems where growth reporting is computed dynamically across a wide score range.

Enabling Computerized Adaptive Testing in large-scale educational assessments under limited connectivity conditions

Gustavo Silva (co-authors: João Augusto Ferreira Freire, Paula Marin Leite Werneck, Jorge de Almeida Carvalho)

A large-scale educational assessment program in Brazil develops and administers assessments for public school systems across different regions of the country. As digital assessment initiatives continue to expand, the program has been exploring the incorporation of Computerized Adaptive Testing (CAT) into its assessment portfolio. However, implementing CAT at scale presents operational challenges associated with the heterogeneous technological infrastructure of participating schools, particularly in contexts where internet connectivity is limited or unstable.

To address this challenge, we propose an architecture integrated with the organization's existing assessment and monitoring platform that minimizes dependence on continuous communication with central servers during test administration. The proposed solution seeks to reconcile two often competing requirements in large-scale adaptive assessment: the need for uninterrupted execution of the adaptive algorithm and the connectivity constraints faced by many participating schools, a challenge that is particularly relevant in large-scale public educational assessment programs in Brazil.

In this architecture, the core CAT components are executed locally on the participant's device, allowing item selection, ability estimation, and session management to take place independently of immediate internet availability. Meanwhile, the platform remains responsible for authentication, assessment administration, session management, and synchronization of assessment data. This approach reduces reliance on continuous network connectivity while preserving the operational processes already established for digital assessment programs. The project is currently in the prototyping and architectural design phase.

An initial CAT engine prototype has been developed in JavaScript to support integration with the organization's assessment and monitoring platform. Preliminary validation using item parameters and ability estimates derived from an operational assessment demonstrated a high degree of numerical consistency in ability estimation between the JavaScript prototype and the R-based routines currently employed by the psychometric team. The observed differences were negligible ($MAE = 3.09 \times 10^{-9}$; $RMSE = 1.39 \times 10^{-8}$). In parallel, an initial integration design has been developed, defining component responsibilities, interaction workflows during test administration, and strategies for local execution in low-connectivity environments.

This presentation discusses the proposed architecture, the preliminary results obtained from the CAT engine prototype, and the main challenges involved in implementing large-scale adaptive assessments in educational settings characterized by heterogeneous technological infrastructure.

Development of a Computerised Adaptive Test for Assessing Upper-Basic School Students' Mathematics-Ability in Nigeria: A Post-hoc Simulation Testing

Isaac Olawale Ifinju

Computerised Adaptive Testing (CAT) is an advance method for educational assessment that has gained popularity in modern assessment practices due to its technology and ability to provide accurate, efficient, and precise assessments. Most assessments in Nigeria are currently done using conventional fixed-form tests, where students, irrespective of their ability, are offered the same test, thus limiting test ability precision, making it inefficient. To address these issues, a Computerised Adaptive Test (CAT) to measure the mathematics ability of Upper-Basic School students was developed using the principles of Item Response Theory (IRT).

The study adopts a research and development (R&D), simulation-based, and evaluation experimental research design involving three phases: the developmental phase, the simulation phase, and the operational deployment field test phase. The developmental phase consists of designing and constructing the system, including an administrative module for item bank management, test configuration, candidate management and efficiency dashboard. Also, an examinee simulation module that allows for an adaptive item administration environment and response generation; and the CAT architecture. The CAT architecture includes a calibrated item bank of 150 Rasch-calibrated mathematics items, a Rasch estimation engine on the logit scale, an ability estimation module based on Maximum Likelihood Estimation (MLE) and Weighted Likelihood Estimation (WLE), a starting rule which selects an item of average difficulty ($b = 0.00$ logit), an adaptive item selection module based on Maximum Fisher Information (MFI), a content balancing module using a constrained version of the Weighted Deviation Model (WDM) and a stopping rule engine which terminates testing when a specified measurement precision ($SEM \leq 0.35$) is met or when a maximum test length of 40 items is reached.

The second phase (which is the focus of this study) is a post hoc simulation testing using the calibrated item bank to produce a set of artificial adaptive test administrations information for a variety of testing conditions for the developed CAT system. The calibrated item bank will be uploaded into the "AllTestSim" software, and 5,000 simulated examinees will be generated. Ability estimates, standard errors, item exposure rates, classification bias, and RMSE will be extracted and analysed using descriptive statistics and ANOVA procedures. This is to determine the testing model that provides the highest measurement precision and efficiency while maintaining acceptable item bank utilisation and score accuracy. Different CAT configurations will be systematically simulated by manipulating item selection methods (MFI, EBI), ability estimation procedures (MLE, EAP, MAP), exposure control strategies, and stopping rules.

The empirical data obtained during the simulation will be used to further refine, optimise and operationally set the developed CAT for subsequent field testing with real Upper-Basic School students. It is expected that the study will give empirical evidence of the psychometric strength and operational readiness of the CAT as well as contribute to the advancement of computerised adaptive assessment within Nigerian basic education.

Designing multi-stage CAT assessments of academic English language proficiency to give examinees the greatest opportunity to demonstrate that they meet performance level thresholds

Stephen Walker

Designing multi-stage CAT assessments of academic English language proficiency to give examinees the greatest opportunity to demonstrate that they meet performance level thresholds.

At the English language teaching center at the University of Queensland, we deliver academic English language programs for international students who use English as an additional language. These students have already met the academic prerequisite requirements for university degree commencement but have not yet met the English language proficiency level specified for entry into their chosen degree courses. The cohorts are largely homogeneous in ability meaning that many students are clustered around the critical proficiency level set as the 'minimum entry' level for Australian degree courses.

Such a truncated proficiency profile, missing both high and low ability learners, poses multiple challenges for test designers. If most students are clustered just at, above or below the minimum proficiency level required, how can assessments differentiate well between the abilities of learners when the cohort itself shows little variation in ability?

In this presentation we outline the development process followed to meet this challenge. We set out our multi-stage CAT design for assessments of both academic reading and listening proficiency which employ a series of polytomous testlet stages each consisting of a long stimulus text and an accompanying set of 6 items dependent on that text. We demonstrate how we avoid possible distortions such as inflated reliability estimates and inflated latent trait precision measurements due to local item dependence between testlet sub-items.

We then show how our use of a Bayesian Maximum Falsification approach to next testlet-item selection permits us to:

- improve the test taker experience by including testlets where we expect near full success rather than the expected 50% success of a maximum information approach
- deliver a test where testlets are selected to continuously challenge the examinee to falsify our existing measure estimate thereby giving each student the maximum opportunity to demonstrate their true performance level

In addition to these test taker benefits, we show how employing this test design in our context can reduce the item exposure that would occur under the statistically optimal maximum information approach in which always selecting the ""best"" item can lead to test-takers of similar ability responding to the same set of targeted items.

References:

Linacre, J.M. Computer-Adaptive Testing, CAT: A Bayesian Approach: Rasch Measurement Transactions, 1995, 9:1 p.412

Lange R. Binary items and beyond: a simulation of computer adaptive testing using the Rasch partial credit model. Journal of Applied Measurement. 2008;9(1):81-104.

Using Semantic Embeddings to Examine Content Overlap Between the IDCP-2 and PID-5

Pedro Godoy dos Santos (co-authors: Leonardo de Barros Mose, Izabella de Araújo Menardi, Ricardo Primi, Lucas de Francisco Carvalho, Giselle Pianowski)

Establishing content validity traditionally relies on expert judgment to evaluate the correspondence between test items and the constructs they are intended to measure. Recent advances in natural language processing provide alternative tools for examining semantic relationships among psychological assessment instruments. This study investigated the use of semantic embeddings as a source of content-validity evidence by examining the correspondence between the Dimensional Clinical Personality Inventory 2 (IDCP-2), a Brazilian dimensional measure of personality pathology, and the Personality Inventory for DSM-5 (PID-5). Semantic embeddings were generated for all 212 IDCP-2 items and all 220 PID-5 items using OpenAI's text-embedding-3-large model via package `embeddcv` (<https://github.com/rprimi/embeddcv>, developed by the authors). Cosine similarity was examined at two levels. First, individual IDCP-2 items were compared with aggregated embeddings representing each of the 25 PID-5 facets. Second, IDCP-2 facet embeddings, computed by averaging the embeddings of items within each facet, were compared with aggregated PID-5 facet embeddings. To identify the strongest semantic correspondences, we adopted an exploratory threshold of two standard deviations above the mean similarity score.

Results indicated substantial semantic convergence between theoretically related constructs across the two instruments. In the item-to-facet analysis, cosine similarities ranged from .06 to .87 ($M = .37$, $SD = .09$), with 200 of 5,300 comparisons exceeding the convergence threshold. In the facet-to-facet analysis, similarities ranged from .14 to .78 ($M = .37$, $SD = .07$), with 53 of 1,175 comparisons exceeding the threshold. Several facets showed strong theoretically expected correspondences, including Eccentric Style with PID-5 Eccentricity (cosine = .78), Emotional Intensity with Emotional Lability (cosine = .77), Intimate Relationships with Intimacy Avoidance (cosine = .74), Risk Taking with Risk Taking (cosine = .70), and Anxiety with Anxiousness (cosine = .69). The analyses also identified facets with meaningful cross-loadings across multiple PID-5 dimensions, suggesting blended dimensional content within some IDCP-2 facets.

At the same time, several findings illustrate important limitations of semantic approaches. In some cases, embeddings produced high similarity between theoretically opposing constructs, indicating that semantic proximity may reflect shared linguistic content rather than clinically meaningful distinctions. Overall, the results suggest that semantic embeddings can provide scalable and informative evidence regarding content overlap between psychological measures, while complementing rather than replacing traditional expert-based content-validity procedures. These findings support the potential utility of embedding-based methods for psychometric research, instrument development, and the systematic comparison of assessment instruments.

AI-Based Item Parameter Estimation as an Alternative to Pilot Testing in Computerized Adaptive Testing: A Simulation Study

Zeus Bellido (co-authors: Alan Mead, Reese Butterfuss)

Developing a computerized adaptive test (CAT) requires calibrated item parameters, traditionally obtained through large-scale pilot testing — a process that is costly, time-consuming, and logistically demanding, particularly for certification programs with limited examinee pools. This requirement limits the application of CAT to many exam programs with lower volumes.

This presentation introduces and evaluates an alternative approach: using large language models (LLMs) to estimate item parameters prior to operational administration by analysis of the content of the item.

The foundation of this work is an empirical validation study conducted across eight professional certification examinations. In each case, an ensemble model was used to estimate item difficulties from the text of the items. Multiple prompts were used to ask the model to estimate the proportion of candidates likely to answer each item correctly and the responses were summarized into a point estimate of item difficulty.

These AI-generated difficulty estimates were then compared to empirically derived proportion-correct statistics from operational administrations. Across the eight exams, AI estimates correlated with empirical values at approximately $r = 0.80$, a level of accuracy that invites serious consideration of AI estimation as a psychometric tool. The presentation will describe the prompting methodology including how item content and a description of the minimally-qualified candidate were communicated to the AI, and will discuss sources of both predictive success and residual error.

Building on this empirical foundation, the core of the presentation is a CAT simulation study comparing two item bank calibration strategies: (1) traditional pilot testing using a modest operational sample, and (2) AI-generated parameter estimates. The simulation adopts realistic certification testing conditions, including 3PL item response models, variable-length stopping rules, and a pass/fail decision framework. Both strategies are evaluated on two key psychometric outcomes: test reliability (squared correlation of true theta and theta-hat) and decision consistency (the proportion of candidates who would receive the same pass/fail outcome across hypothetical replications).

Results are expected to demonstrate that AI-estimated parameters, despite their inherent imprecision, can support CAT administrations with reliability and decision consistency approaching those achieved with conventionally piloted item banks based on modest sample sizes. We will discuss the conditions under which AI estimation is most and least adequate will be discussed.

This research has direct practical implications for certification bodies seeking to launch or expand CAT programs without the infrastructure for large-scale pretesting. It also opens a methodological conversation about the appropriate role of AI in the psychometric pipeline — not as a replacement for empirical data, but as a principled, scalable source of prior information.

Adaptive Testing Meets Contracting: Rethinking Success in the Age of AI

Xianguen Hu (co-author: Duanli Yan)

Computerized Adaptive Testing (CAT) has matured into a highly efficient and psychometrically grounded approach to assessment, dynamically selecting items to estimate examinee ability with minimal test length while maintaining precision. Traditionally, CAT is conceptualized around a fixed achievement objective: every examinee progresses adaptively toward a common endpoint, with individual differences in prior knowledge and skill informing item selection but not altering the ultimate standard of success. However, this primacy of fixed end-state performance is increasingly challenged in the era of AI. In response, this paper proposes a conceptual and practical shift toward goal-adaptivity through contract-based assessment, operationalized in a novel adaptive testing system inspired by contracting bridge games.

In contracting bridge, players declare a contract, a conditional achievement goal based on their hand, and success is judged by fulfillment of that contract rather than by maximizing absolute outcomes. This logic foregrounds human competencies that remain highly relevant in the AI era, including realistic self-assessment, strategic calibration of ambition, commitment under uncertainty, and evidence-based adjustment of goals. Translating this framework into adaptive assessment, we reconceptualize adaptivity as involving not only adaptive item selection, but also adaptive goals negotiated according to an examinee's starting point and declared intent.

To operationalize this approach, we developed an assessment system in which cards represent calibrated task levels, and examinees declare a contractual target consisting of the levels or sequence of tasks they aim to complete successfully. The declaration reflects a commitment grounded in self-perceived capability rather than a universal proficiency benchmark. Once a contract is established, the adaptive engine dynamically selects and sequences tasks aligned with both the declared ambition and observed performance. Unlike traditional CAT, where termination is based on psychometric criteria such as standard error or maximum information, success in this system is determined by whether the examinee fulfills the declared contract under adaptive conditions.

The implementation includes a computational backend that maps cards to hierarchical task clusters with calibrated difficulty analogs, an adaptive selection strategy that balances challenge and support within the declared contract range, and metadata logging that captures bid calibration accuracy, adjustment strategies, and in-contract learning gains. The resulting assessment profile extends beyond latent trait estimation to include measures of goal calibration, strategic adjustment, and growth relative to targeted aspiration, aligning with current interests in formative assessment and competency-based measurement.

This work contributes both a theoretical reframing of adaptivity, from fixed outcome optimization to conditional goal fulfillment, and a functioning implementation suitable for empirical study within CAT research. It also aligns with contemporary interests in AI-informed assessment innovation by placing human agency and strategic self-regulation at the center of adaptive measurement.

AI-Assisted Cognitive Pretesting for Personality Assessment Items: Comparing Profile-Conditioned Simulated Responses With Observed Human Data

Diego Forteza (co-authors: Luke Treglown, Reece Akhtar)

Large language models have been proposed as tools for simulating human samples and respondent behavior (Aher et al., 2023; Argyle et al., 2023), but their role in psychometric development remains unsettled. Consistent with validity frameworks that distinguish item-review evidence from operational score validation (AERA, APA, & NCME, 2014), this presentation evaluates a cautious workflow for using AI-generated responses as a diagnostic supplement to expert item review and cognitive pretesting (Willis, 2005) in a non-cognitive personality assessment. Rather than treating simulated responses as substitutes for human validation data, the study asks whether empirically grounded prompts can reproduce selected response patterns closely enough to prioritize items for qualitative review.

Using observed responses from 100 English-locale respondents to the six-dimensional Core Drivers assessment, we evaluated 90 scored items across Agreeableness, Conscientiousness, Drive, Extraversion, Openness, and Stability. Two prompting strategies were compared in a held-out reconstruction task: profile-conditioned prompting, in which the model received dimension and subfacet profile information, and response-history prompting, in which the model received prior observed item responses. The profile-conditioned approach showed stronger reconstruction performance, with lower ordered-item mean absolute error (0.62 vs. 0.81), higher within-one agreement (92.5% vs. 82.2%), higher forced-choice exact-match accuracy (78.2% vs. 70.5%), and higher mean item-level distributional correlation (0.50 vs. 0.32).

A second analysis generated profile-conditioned responses for the broader item bank and compared simulated and observed response matrices using response-level reconstruction, classical item and dimension summaries, Rasch partial-credit diagnostics (Masters, 1982), and a source-based item-functioning screen adapted from DIF logic. Among evaluable pairs, forced-choice exact-match accuracy was 76.5%; opinion-scale responses had mean absolute error of 0.59 and within-one agreement of 91.6%. At the same time, simulated responses diverged from observed data in several psychometrically important ways: response coverage was lower, mean scores were higher across all six dimensions, Rasch EAP reliability was lower, item-location recovery was limited, and 26 items were flagged for source-based item review.

These findings suggest that AI-generated response simulation may be useful as a structured item-review triage layer before larger-scale calibration or field testing, but should not be interpreted as operational calibration or validation evidence (AERA, APA, & NCME, 2014). The presentation offers a workflow for separating potentially useful item-review signals from artifacts of prompt design, missingness, and model behavior.

Use of AI to Assess Metaphor Generation: Convergent And External Validity

Beatriz Chrispim (co-authors: Izabella de Araújo Menardi, Pedro Godoy dos Santos, Natalia de Oliveira Santos, Cibelle de Oliveira, Felipe Feifar Afonso, Carolina Pupo Nogueira Cervone, Rosiane Gonçalves Coelho Silva, Ricardo Primi)

The creation of metaphors is a complex form of thought associated with creativity and the personality trait Openness to Experience. This study examined evidence of validity for scores produced by artificial intelligence (AI) in the Teste de Criação de Metáforas (TCM), comparing AI-based scoring with human expert ratings. We also investigated whether AI-generated scores showed external validity patterns comparable to those obtained from human raters, through correlations with Openness, creative achievement measured by the Kaufman Domains of Creativity Scale (K-DOCS), and group indicators related to musical expertise. The sample consisted of 131 adults, with a mean age of 28 years, of whom 85 were female. Responses were scored by four human experts and by several AI models: GPT-4o, GPT-4o-mini, GPT-4.1, DeepSeek using few-shot learning, and BERTimbau using fine-tuning. Results indicated fair to moderate agreement between human experts and generative AI models, with kappas ranging approximately from .17 to .46 and Pearson correlations from .19 to .48. These values are comparable to human–human agreement found in this sample. Agreement among generative AI models was higher, with correlations ranging from .48 to .78. BERT-based models showed weaker convergence with both human raters and generative models, although the two BERT specifications were highly consistent with each other. A Many-Facet Rasch Model (MFRM) indicated adequate fit for most items and raters, with infit and outfit values generally within acceptable ranges. The model also revealed differences in severity between human raters and AI systems, with rater severity estimates ranging from -0.92 to 0.81 logits, indicating that some AI models behaved similarly to lenient human raters whereas others were relatively more severe. External validity evidence supported the use of generative AI scores. Correlations with Openness were positive and comparable to, or slightly higher than, those obtained with human raters. Similar patterns were observed for creative achievement indicators, especially general and artistic/academic creativity, for which generative AI models showed small to moderate validity coefficients. In contrast, BERT-based models showed weaker and less consistent external validity. Overall, the findings suggest that generative AI models can approximate human judgment in the assessment of metaphor creation and produce validity evidence comparable to expert ratings. At the same time, the results indicate that AI architectures differ substantially in their capacity to reproduce human evaluations of creativity. These findings support the potential of AI-assisted scoring to increase scalability and efficiency in the assessment of metaphor creation, provided that model-specific validity evidence and human oversight are maintained.

Psychometrics Meets LLM: Diagnostic Assessment Copilot Design and Deployment

Chun Wang

Research consistently shows that providing personalized feedback and tailored learning guides can help improve learning. However, this level of personalization is often not possible in high-enrollment classes. To fill this gap, we developed a general purpose “diagnostic assessment copilot,” a web-based application with an instructor-facing interface to generate and catalog assessment items. This copilot uses large language model (LLM) to automate the two most resource-intensive components of diagnostic assessment design: item generation and item-to-facet mapping. These tasks traditionally require iterative, intensive collaboration among content experts, learning scientists, and psychometricians, making high-quality diagnostic assessments prohibitively expensive for most courses. By leveraging LLM to generate contextually appropriate content while maintaining alignment to defined knowledge facets through a human-in-the-loop workflow, the copilot unlocks assessment-as-learning at scale.

The design of this copilot tool is grounded in the learning science theory of conceptual change. The theory advocates that the key to targeted instruction is accurate and reliable evaluation of students’ understanding, which includes both correct understanding, denoted as goal facets, and misconceptions, denoted as intermediate facets. The copilot contains two modules: (1) Item generation module: it can automatically generate multiple choice (MC) items with diagnostically rich distractors so that each MC item is capable of measuring a defined combination of goal facet(s) and coherently related misconceptions, and (2) Item facet mapping module: LLM will be used as a judge to evaluate how well those auto-generated items assess the intended construct, thereby providing feedback to the item generation module.

In this talk, I will introduce our general-purpose diagnostic assessment copilot as well as the diagnostic facet status model (DFSM) and neural network cognitive diagnostic models (neuralCDM), two types of CDMs that can score the assessment results and output students’ cognitive profiles. I will also introduce our latest studies that use LLM output, augmented with students’ responses, to refine Q-matrix. Lastly, I will describe the adaptive item selection methods for learning with CDMs.

Automated Evaluation of Teacher Portfolios Using Large Language Models and Structured Educational Evidence Analysis

Gustavo Silva (co-authors: João Augusto Ferreira Freire, Jorge de Almeida Carvalho)

Teacher evaluation serves as a cornerstone for enhancing instructional quality and supporting continuous professional development within modern educational systems. When implemented formatively, this process offers tailored feedback that encourages educators to reflect deeply on their lesson planning, classroom strategies, and curriculum alignment. However, executing meaningful teacher evaluations at scale presents a monumental operational bottleneck. In Brazil, for instance, data from the 2024 School Census indicates that there are approximately 2.4 million basic education teachers actively working in the school system. Managing, reviewing, and providing qualitative, pedagogically sound feedback for millions of unique teacher portfolios manually is not only incredibly time-consuming but also financially prohibitive for public educational institutions.

To address this scalability crisis without sacrificing the individualized nature of formative assessment, artificial intelligence offers a viable path forward. Large Language Models have demonstrated a remarkable capacity for text analysis, reasoning, and context synthesis. When combined with advanced information retrieval techniques like Retrieval-Augmented Generation, these models can move beyond generic text generation. They can actively ground their assessments in official regulatory frameworks, map pedagogical evidence to explicit standards, and generate actionable insights for classroom teachers.

This study introduces an automated pipeline designed to evaluate teacher portfolios by combining Large Language Models with a specialized retrieval component. The primary objective is to process written reflections and submitted pedagogical evidence, verify their alignment with educational rubrics, and deliver consistent, individualized feedback. Central to this system architecture is the integration of Brazil's National Common Curricular Base. The text of this massive framework was meticulously preprocessed, breaking down its extensive skill sets into discrete descriptors. These descriptions were enhanced via language models and indexed into a FAISS vector database to optimize data retrieval accuracy.

The system operates by extracting abstract learning descriptors from a teacher's written report, querying the vector database to retrieve the most relevant national skills, and passing this context to the primary language model. The model then weighs the teacher's input and supporting documents against established evaluation rubrics to construct a comprehensive pedagogical report. To evaluate the practical efficacy of this pipeline, the study utilizes a dataset of 200 authentic teacher portfolios from the Minas Gerais State Education Network. Given the subjective and nuanced nature of educational assessment, these AI-generated reports are validated by human educational specialists who score them for quality, relevance, and pedagogical utility.

Ultimately, this research demonstrates the viability of utilizing cutting-edge natural language processing to alleviate the operational burdens plaguing large-scale educational systems. By anchoring the model's intelligence to official state frameworks, the pipeline mitigates common artificial intelligence limitations such as hallucinations, ensuring that the automated feedback remains deeply relevant and trustworthy. This framework establishes a foundational model for scalable, cost-effective, and deeply personalized

professional development tools, mapping a clear path from regional testing in Minas Gerais toward a potential nationwide deployment.

Methodology for the Empirical Generation and Validation of Performance Level Descriptors Using Generative Artificial Intelligence

Daniela Jimenez (co-authors: Paulina Barriga, Camila Straub)

This paper evaluates whether general-purpose generative AI can support the generation and validation of Performance Level Descriptors (PLDs) in large-scale assessment, offering a reproducible, low-cost method for collecting validity evidence. The approach is applied to the Mathematical Competence 1 (M1) section of Chile's Higher Education Admission Test (PAES), used in the country's centralized national admissions system, building on performance levels developed in a prior research project and validated by expert judgment.

The evaluation is organized around a three-phase workflow: (i) empirical construction of clusters and descriptions of performance levels from items and Item Response Theory (IRT) difficulty parameters; (ii) validation that each level's descriptor coheres with the evidence from items assigned to it, including blind classification tests assessing the mutual exclusivity of adjacent levels; and (iii) alignment between items from recent operational applications and the descriptor scale, to identify regions of the latent variable where adjacent levels are difficult to distinguish or where item density is low.

Unlike the development of custom technological solutions or complex software architectures, this approach facilitates access to workflows using readily available AI tools, allowing institutions and technical committees to implement reviews of performance level descriptions without requiring programming skills or specialized infrastructure.

Preliminary results indicate that the implementation of this approach significantly reduces development time, optimizing the cycle from system prompt planning to the final analysis of results. In this schema, the team's effort shifts from extensive manual item classification processes and drafting descriptions to the preparation and systematization of input data. This involves properly structuring the item bank, its characterization, and its statistical metadata (such as the difficulty parameter). Results so far indicate that the LLM also generates accurate semantic syntheses that characterize the different clusters. We conclude discussing the potentially complementary role for human expert judgment and general-purpose LLMs in PLD development, providing a viable route to validity evidence for score interpretation in large-scale assessment.

A Cognitive Diagnosis Model for Latent Classification of Bounded Continuous Variables

Eduardo Schneider Bueno de Oliveira (co-authors: Xiaojing Wang, Jorge L. Bazán, Jimmy de la Torre)

Cognitive Diagnosis Models (CDMs) are widely used in latent-variable modeling for classification tasks that diagnose abilities, skills, or attributes. Originally developed for dichotomous indicators, CDMs have been extended to polytomous and continuous responses, including bounded continuous variables such as proportions and index scores. In this presentation, we introduce a Bounded DINA (B-DINA) model, an extension of the DINA model for handling bounded continuous responses using a Beta distribution with a mean–precision parameterization. We present a Bayesian estimation framework, define interpretable item parameters, and compute posterior probabilities of membership in each latent-attribute profile. The proposed methodology explicitly addresses label-switching nonidentifiability and assesses model fit through posterior predictive checks.

To evaluate the performance of the proposed approach, we conduct a simulation study examining parameter recovery under different sample sizes, test lengths, and levels of item discrimination. The results indicate satisfactory recovery of both item and person parameters, with improved performance as sample size and item discrimination increase. We further illustrate the model using municipal educational, social, and economic indicators from Southeastern Brazil. The analysis provides probabilistic classification of municipalities into latent profiles and reveals relationships between observed indicators and latent attributes. As bounded continuous variables are common across educational, psychological, social, health, and policy-related research, the proposed B-DINA model provides a framework for latent classification based on bounded responses.

Black-Box Variational Inference with REINFORCE for the DINA/DINO Model

Flavio Barros

Cognitive diagnosis models (CDMs), such as DINA and DINO, represent respondents through binary latent attribute profiles. A major computational difficulty arises from marginal-likelihood inference, which requires summation over the 2^K possible latent profiles and therefore becomes increasingly expensive as the number of attributes K grows.

This work presents a non-amortised black-box variational inference (BBVI) approach for the DINA and DINO models based on the score-function estimator (also known as REINFORCE). For each respondent, the approximate posterior distribution is factorised into K Bernoulli components with local variational parameters. Binary attribute profiles are sampled directly from this distribution, ensuring that the DINA and DINO ideal-response functions remain discrete throughout optimisation. Gradients with respect to the variational parameters are estimated through the score-function identity, while the item parameters are optimised by automatic differentiation.

The method was evaluated using 13 simulation scenarios covering $K = 6-40$, $J = 40300$, and up to $N = 20_000$ respondents, with 100 replications per scenario. The estimator remains operational in the high-dimensional stress scenario with $K = 40$, $J = 300$, and $N = 20_000$. Marginal attribute recovery remains high across the investigated dimensional range, whereas exact recovery of the complete attribute profile becomes progressively more difficult as K increases. These results show that score-function BBVI provides a direct way of performing variational inference for high-dimensional cognitive diagnosis models while preserving discrete latent attributes and avoiding explicit enumeration of the 2^K latent state space.

Automatic Pre-Testing of Mathematics Assessment Items: Predicting Item Response Theory (IRT) Difficulty Parameter with Machine Learning Methods

Patrick Canto de Carvalho (co-authors: João Augusto Ferreira Freire, Luan Henrique da Silva Barbosa, Emerson Caneschi Coelho de Souza)

The calibration of new items in educational assessment constitute a fundamental but costly process. Item Response Theory is widely used to measure latent traits, such as student proficiency, through parameters that represent items' traits: discrimination, difficulty, and guessing. Nonetheless, the need to pretest new items at scale to provide parameters comes with high costs, and limits subject matter experts if adjustments are required - which would entail new rounds of pretesting and further calibration. In this context, predicting item parameters could improve the efficiency of large-scale assessments by reducing the reliance on resource-intensive pretesting procedures and subsequent recalibration efforts.

We propose to tackle the problem of automatically estimating the difficulty parameter of new items using item text and metadata for additional information from previous exams through natural language processing and supervised machine learning techniques, with data of mathematics assessment items administered to students in basic education (grades 1–12) in Brazil.

This study investigated the feasibility of predicting Item Response Theory (IRT) difficulty parameters and assesses their potential to support item calibration and pretesting directly from item textual content, metadata, machine learning techniques, evaluating through the process multiple regression models using different text representations, including Bag-of-Words, TF-IDF, Word2Vec, and ModernBERT embeddings. We used a tokenizer that converts characters into standardized tokens, and Large Language Models (LLMs) to extract features from item images that share contexts or serve as support.

For this study, the data used for mathematics assessment items were divided into two subsets: grades 1–3, with 2720 items, and another comprising grades 4–12, with 8155 items each one on different scales and analyzed separately. Model performance was assessed using repeated cross-validation on the training set and a held-out test set, while paired Wilcoxon tests were used to assess the statistical significance of performance differences.

And as results for the grades 1–3, Random Forest with TF-IDF, achieved an MAE of 0.4734 and an RMSE of 0.6876, and showed difficulty with edge cases—likely due to its significantly smaller dataset. For grades 4–12, Random Forest with a Bag-of-Words (BoW) representation yielded the lowest errors (MAE: 0.4957, RMSE: 0.7040), performs better overall but still struggles with higher b parameter items.

Results show that the proposed models achieve consistently lower mean absolute error in predicting IRT difficulty parameters than previously reported approaches, demonstrating improvement upon those previously reported results in the literature. The findings indicate that even with prediction errors in the difficulty parameter, the impact on students' final proficiency estimates remains substantially smaller than the standard deviation of the proficiency scale, suggesting that prediction inaccuracies may have a limited impact on downstream ability estimation. Future research should leverage Large Language Models to better capture the cognitive demands of mathematical problem-solving and to simulate examinee responses. Additionally, upcoming analyses must

evaluate the importance of specific skills, features, and distractor characteristics as predictive signals.

Accelerating Item Piloting with NLP-Based Features and Bayesian CAT

Steven Nydick (co-authors: Manqian Liao, J.R. Lockwood, James Sharpnack, Alexander Tsigler)

Large-scale digital assessments rely on large item banks to maintain measurement precision, content coverage, and test security. While recent advances in automatic item generation, such as those based on large language models, have allowed test developers to quickly expand item banks, traditional piloting and calibration practices remain a bottleneck. Newly generated items must typically be piloted as unscored and randomly administered, resulting in inefficient test delivery and delayed integration into operational pools. To address those inefficiencies, we propose a unified framework for accelerating item piloting by incorporating natural language processing (NLP)-based item features into a Bayesian CAT system, enabling the operational use of items with substantial item parameter uncertainty.

The proposed system consists of three components: Bayesian Item Parameter Calibration, Soft Scoring (S2), and Adaptive-Adaptive Administration (A3). First, NLP-based features are extracted from item content. These features, combined with available response data, are included into a Bayesian calibration model that produces (posterior) samples of item parameters. This model can then be used to obtain estimates of item parameters for newly written items along with corresponding uncertainty based entirely on their features without response data.

The S2 component estimates ability by integrating over item parameter uncertainty. Rather than conditioning on point estimates of item parameters (such as the posterior mean or mode), the likelihood contribution of an item response is averaged over the posterior samples of item parameters. This “soft” scoring approach results in downweights to items with high uncertainty while still including them as scored. Thus, newly generated items (with just features) or items with limited responses can still be included in scoring.

The A3 component modifies adaptive item selection to account for item parameter uncertainty. Rather than defining a single selection criterion (such as Fisher information computed at fixed parameters) A3 introduces an additional layer of stochasticity by sampling from the posterior distributions during selection. Therefore, items with more uncertainty in their item parameters will be administered to a wider range of test takers, facilitating data collection for subsequent calibration.

Together, these components enable continuous piloting. In the S2A3 system, items transition from high uncertainty, low weights, more exploration in administration to stable, high weights and more targeted administration. Empirical results demonstrate that NLP-based item features meaningfully predict item difficulty and that S2A3 substantially reduces the number of responses required to reach acceptable parameter precision. We will provide examples based on the current administration of the Duolingo English Test (DET), highlighting both practical implementation and implications for large-scale adaptive assessment systems.

LLM and Rule-Based Automatic Item Generation: A Two-Study Investigation in Large-Scale Literacy Assessment

Lucas Larcher (co-author: Jairo Francisco de Souza)

Developing high-quality assessment items for large-scale assessments is a costly and specialized process requiring deep curricular knowledge, formal training in item writing, and strict adherence to psychometric standards. Although generative AI has been widely proposed as a tool to support automatic item generation (AIG), empirical evidence on its effectiveness under domain-specific structural constraints remains limited — particularly when tasks impose simultaneous linguistic, pedagogical, and psychometric requirements that go beyond textual fluency. This paper presents two complementary studies examining the role of LLM-based and rule-based approaches in AIG, under different types of task constraints, in the context of Brazilian large-scale literacy assessments targeting early elementary students.

The first study examines whether a multi-agent pipeline can produce Portuguese literacy items assessing the ability to identify logical-discursive relations signaled by conjunctions — a skill requiring discourse-level reasoning rather than surface-level pattern matching. Three specialized agents handle generation, evaluation, and iterative revision, guided by structured prompt templates derived from official item-construction guidelines developed by assessment specialists. Three large language models were compared across multiple prompt engineering strategies. Expert evaluation revealed that distractor quality was the primary and persistent bottleneck across all conditions. The multi-agent architecture yielded meaningful improvements over single-prompt generation for lower-complexity items, but performance degraded substantially under higher cognitive and linguistic demands. Crucially, prompt engineering strategies produced no consistent quality gains, suggesting that item quality in structured AIG tasks is governed more by task complexity than by prompting technique. These findings challenge the assumption that architectural sophistication alone is sufficient to meet the validity and guideline-adherence standards required in operational assessments.

The second study investigates a more constrained problem: generating distractor candidates for word-reading items, where alternatives must simultaneously satisfy phonological, graphemic, syllabic, and pedagogical criteria. We argue that tasks governed by explicit formal rules constitute a distinct class of AIG problems for which LLM-based approaches are structurally ill-suited. To test this hypothesis, we developed a rule-based linguistic approach supported by a structured lexical database annotated with syllabic, phonological, and frequency information. Candidates are filtered by structural constraints and ranked by a similarity function operating at the syllabic level. Evaluated against few-shot LLM strategies across both quantitative distractor recovery and blind specialist assessment, the rule-based approach consistently outperformed LLMs, with a statistically significant advantage for non-canonical words — the most psychometrically demanding profile. LLM-generated lists produced a substantially higher rate of invalid or nonexistent forms, particularly for structurally complex items.

Taken together, these studies advance a principled distinction between AIG tasks that benefit from the contextual flexibility of LLMs and those that require deterministic adherence to formal constraints. We propose that effective AIG pipelines should treat these as complementary rather than competing paradigms: LLMs for drafting and contextual judgment, rule-based systems for constraint verification and candidate

generation. Both studies also converge on a shared finding: human expert oversight remains a necessary condition for validity in operational assessment contexts, not merely a transitional safeguard.

An IRT-Base Adaptive Measure Of Conscientiousness In The Big Five Framework

Gabriela S. Lozzia (co-authors: Facundo Juan Pablo Abal, Horacio Félix Attorresi, María Silvia Galibert)

Within the Big Five Model, Conscientiousness is a personality domain that describes individual differences in self-regulated behavior, encompassing both inhibitory and proactive aspects of conduct. This study presents the results of the development of an adaptive test for the assessment of this domain using Item Response Theory (IRT). A total of 934 adults from the general population of Buenos Aires, Argentina participated in the study (50.8% female; mean age = 35.3 years, SD = 14.1, range = 18–85). Participants completed a 24-item instrument with four response options designed to measure Conscientiousness according to the definition proposed by McCrae and Costa (2010). Expert judges' agreement regarding item content was examined using Aiken's coefficient. A randomly selected 70% subsample was used to calibrate the items under the Graded Response Model using IRTPRO, while the remaining 30% was used to evaluate adaptive measurement performance through FIRESTAR simulations. Items showed adequate fit to the polytomous model, and no differential item functioning by sex was detected using the Wald test. To evaluate adaptive administration, an algorithm was programmed using the trait mean as the starting criterion. Items were progressively selected according to Fisher's Maximum Information method. Two stopping criteria were compared: (a) achieving a standard error $\leq .30$ (variable-length administration, approximately equivalent to a classical reliability of .91), and (b) administering a fixed length of 10 items. Under both conditions, adaptive measurement showed limitations in the representativeness of sampled content; however, estimates correlated strongly with scores obtained from the full test ($r = .95$). Relationships between the domain and external criteria were also preserved. Under the variable-length condition, participants required an average of 10.01 items (SD = 4.65, range = 5–24). After 10 administered items, 67.9% of participants achieved a standard error $\leq .30$, and 91.1% required 16 items or fewer. Only 5% of participants failed to reach the predefined precision level before the maximum limit of 24 items was reached. These individuals showed θ scores more than 1.5 standard deviations below the trait mean, although their standard error did not exceed .43 (approximately equivalent to a classical reliability of .82). The mean estimated trait level was -0.25 (SD = 0.84, range = -2.73 to 2.13). Estimation errors ranged from .23 to .43, with a mean of .29 (SD = 0.02). Under the fixed-length condition (10 adaptively administered items), the mean estimated trait level was -0.23 (SD = 0.86, range = -2.81 to 2.01). The mean standard error was .29 (SD = 0.05, range = .17–.48), corresponding to an average classical reliability of approximately .92 and a minimum reliability of .77. Furthermore, 96% of participants were assessed with a standard error $\leq .40$ (approximately equivalent to a classical reliability of .88). These findings indicate that administering fewer than 50% of the original items produced only a slightly higher estimation error compared to the full-length version (.29 vs. .23). Overall, adaptive measurement proved to be efficient, although expanding the item bank is recommended to further optimize measurement precision.

Adaptive Test of Vocational Interests (TIPA): Development and psychometric properties of a computerized assessment measure

Gustavo Martins (co-authors: Ana Paula Salvador, Rodolfo Augusto Matteo Ambiel, Ana Paula Porto Noronha)

The assessment of vocational interests is an essential tool for promoting clients' self-knowledge in career counseling contexts. However, traditional fixed-length interest inventories often require lengthy administrations and may include items that provide limited measurement precision for individuals with specific profiles. These limitations raise an important question: can Computerized Adaptive Testing (CAT) improve the efficiency and precision of vocational interest measures without compromising validity? CAT represents a promising approach because it dynamically selects items based on respondents' latent trait estimates, potentially reducing test length while maintaining—or even improving—measurement quality. To address this gap, the present study developed the Adaptive Test of Vocational Interests (TIPA), using a CAT framework to assess basic interests within the RIASEC model, and examined validity evidence based on internal structure and relations with external variables. Study 1 examined the internal structure of the linear version of the TIPA, implemented the CAT algorithm, and investigated validity evidence based on relations with other variables. The sample included 1,374 participants who completed the TIPA, a sociodemographic questionnaire, and the 18REST-2. Data were collected online, and each participant responded to 160 TIPA items following a balanced incomplete block design adopted in this study. The results of Study 1 supported a hierarchical factor structure for the TIPA, composed of 22 basic interests organized into six second-order RIASEC types. The use of stopping rules based on a standard error of measurement (SE) below .30, a reduction in SE below .015, or a change in theta below .03 resulted in average reductions in test length greater than 50% in simulation study. Despite this reduction, scores from the CAT version were highly correlated with scores from the linear version ($r = .97$). Strong positive correlations were found between the TIPA and the RIASEC types assessed by the 18REST-2. In Study 2, the TIPA was tested in CAT format with an independent sample and in a practical application context. The sample consisted of 231 Brazilians from all five regions of the country, mostly women ($n = 161$; 69.70%), aged between 14 and 62 years ($M = 22.70$; $SD = 9.49$). The TIPA item bank comprises 209 items, and in this sample the number of administered items ranged from 122 to 209 ($M = 168.09$; $SD = 17.67$). Taken together, these findings provide a clear and direct answer to the central research question. The CAT version of the TIPA substantially improved assessment efficiency by reducing test length under multiple stopping criteria, while maintaining strong score similarity with the linear version. These results demonstrate that CAT can significantly enhance the efficiency of vocational interest assessment without loss of precision or validity. The positive correlations with external measures, namely the 18REST-2, consistent with previous literature, further support the validity of the scores generated by the CAT version. Overall, the findings show that CAT is not only feasible but also advantageous for measuring vocational interests, offering faster administrations, reduced respondent burden, and robust score interpretation.

Using Item Response Theory to Develop a Parsimonious Measure of Psychological Need Satisfaction at Work: Implications for Decent Work Research and Social Justice

Alexsandro Luiz de Andrade

Measuring psychological need satisfaction at work in vulnerable and hard-to-reach populations poses persistent methodological challenges. Instruments validated in high-income Western contexts frequently carry item redundancy and respondent burden that compromise data quality precisely where measurement matters most — among workers facing economic constraints, informal employment, and structural marginalization. This study applies Item Response Theory (IRT) to develop and validate two reduced versions of the Work Needs Satisfaction Scales (WNSS; Autin et al., 2019), a 20-item instrument grounded in Psychology of Working Theory (PWT), with a Brazilian sample of 1,470 workers. A Graded Response Model was estimated across all 20 original items. All items demonstrated adequate discrimination parameters ($a \geq 1.60$), confirming that the full-scale functions well psychometrically. A weighted composite criterion combining discrimination, information, and theoretical representativeness guided item selection, yielding a 15-item multidimensional version and a 5-item unidimensional version. Confirmatory Factor Analysis with WLSMV estimator validated both versions: the five-factor structure remained intact after a 25% item reduction ($CFI = 0.995$, $TLI = 0.994$, $RMSEA = 0.088$), and the unidimensional model showed excellent fit ($CFI = 0.996$). Internal consistency was satisfactory across all dimensions ($\alpha = 0.86\text{--}0.91$; $\omega = 0.86\text{--}0.91$), comparable to previous international validations in the United States, South Korea, China, and among Latinx workers. IRT-based item selection produced instruments that are shorter without sacrificing construct coverage. The 15-item version preserves dimensional differentiation across the five need domains: survival, social contribution, competence, relatedness, and autonomy — and is suitable for research requiring theoretical specificity. The 5-item version functions as a parsimonious screening tool appropriate for longitudinal designs, large survey batteries, and clinical monitoring contexts. Critically, inter-version correlations between equivalent dimensions were near-perfect ($r = 0.98\text{--}0.99$), confirming measurement equivalence across versions. From a social justice standpoint, reducing respondent burden without compromising psychometric integrity is not merely a technical advantage. In the Brazilian context, where 70% of the sample earns below six-monthly minimum wages and 21% is informally employed, shorter valid instruments enable systematic monitoring of decent work access in populations that are simultaneously most affected by labor precarity and least accessible through standard research designs. IRT offers a principled, empirically grounded path to this goal: item information functions identify which items contribute most to measurement precision across the latent trait range relevant to vulnerable workers, rather than retaining items based on convenience or tradition. This presentation discusses the methodological decisions, results, and implications for researchers and practitioners working at the intersection of vocational psychology, social justice, and psychometric development in Global South contexts.

Review and Requalification of a Workforce Competency Assessment Tool for Technical Education

Mayara Rachid (co-authors: Fernando de Lollo, Wilson Conciani, Luccas Simão)

Since 2022, a General Competencies for the World of Work (CGMT) Assessment has been administered to public school students in the states of Rio Grande do Sul, Paraná, Minas Gerais, Paraíba and São Paulo. Developed in line with the General Competencies outlined in the National Common Curricular Base (BNCC), the tool seeks to evaluate high school students' and graduates' degree of readiness for entry, retention, and development in the labor market, using a diagnostic instrument designed to measure transversal competencies related to professional pathways and personal development.

Developed by Itaú Educação e Trabalho, a superintendency of the Itaú Foundation, in partnership with the Laboratory for Studies and Research in Social Economics (Lepes) at the University of São Paulo (USP) in Ribeirão Preto, the assessment is structured around three core dimensions: maturity, employability, and purpose. These dimensions aim to capture aspects related to readiness for entering the workforce, adaptability to different work contexts, career planning, and the mobilization of socio-emotional and interpersonal skills within productive environments.

The instrument has already been administered to more than 100,000 students over recent years, establishing a comprehensive database and demonstrating the tool's potential to inform public policy and debates regarding the quality of Vocational Education and Training (VET). At the same time, expanding the scale of implementation and consolidating the response database necessitate a methodological review of the assessment, particularly regarding the instrument's dimensional structure, item design, and potential biases associated with self-report measures—such as social stereotypes and unconscious biases in student self-assessments.

In this context, the updating of the tool proposes its refinement based on new theoretical frameworks regarding youth, future-of-work skills, and non-cognitive skill assessment. The proposal includes reviewing the dimensions evaluated, performing psychometric analysis on the instrument's structure, exploring strategies for administration and result interpretation, and engaging with national and international experts in educational assessment, psychometrics, youth, and labor.

Beyond its diagnostic potential, the tool holds strategic relevance for the contemporary debate on quality in EPT, especially in light of the implementation of the National System for Evaluating Professional and Technological Education (SINAEP) in Brazil. In this scenario, assessing General Competencies can help integrate dimensions related to student development and workforce readiness into educational evaluation systems. Thus, this work aims to present the revision and enhancement process of the CGMT, as well as to discuss the methodological challenges of large-scale measurement of general competencies and their contributions to public education and labor policies.

Design and Implementation of Practice-Oriented, Large-Scale Certification Exams: The Case of ANBIMA's Distribution Certifications

Andrea Burgos de Azevedo Mangabeira

This paper presents the design and implementation of ANBIMA's new certification model for professionals engaged in the distribution of investment products in Brazil. Already implemented on a nationwide scale, the model represents a significant shift from traditional knowledge-based assessments to a practice-oriented approach grounded in professional profiles and their daily professional activities. Certifications such as CPA, C-Pro R, and C-Pro I are structured to reflect distinct professional roles, including client service, relationship management, and technical advisory, promoting strong alignment between certification requirements and market practices.

In this context, ANBIMA certifications go beyond educational credentials, as they are mandatory requirements for professionals to perform distribution activities within institutions adhering to ANBIMA's codes. As such, they function as a prerequisite for professional practice, increasing the expectations of the financial market regarding the reliability, rigor, and credibility of the examination process. This regulatory dimension reinforces the need for a robust and defensible assessment model capable of supporting high-stakes decision-making.

The primary objective of this work is to present and discuss the assessment model developed under this framework, highlighting its conceptual foundations and implementation challenges. A central principle of the model is the direct connection between exam content and the daily activities performed by professionals. Item development is guided by the requirement that questions be contextualized in realistic scenarios, where candidates must apply knowledge in situations that simulate professional decision-making. The item development guidelines explicitly emphasize that question design must consider the activities and responsibilities associated with each certification profile, ensuring that the assessment accurately reflects professional practice.

Despite its conceptual strengths, implementing this model at scale introduces significant challenges. ANBIMA's exams are delivered digitally, on a continuous basis, and across the entire Brazilian territory, requiring high levels of standardization, consistency, and operational robustness. Ensuring comparability across candidates while preserving contextualization is a key complexity in this large-scale environment.

Another critical challenge lies in the development of interactive item formats. These models aim to move beyond traditional multiple-choice questions, incorporating structures that better simulate decision-making processes. However, such formats must maintain a high level of rigor, allow for fully automated scoring without human intervention, and be designed in a way that minimizes vulnerabilities to fraud. Certification regulations reinforce the importance of preventing irregularities throughout the process, further increasing the demands placed on exam design and security.

In summary, the implemented model represents an integrated approach that combines practice-oriented assessment, technological scalability, and rigorous security standards. This paper contributes by documenting the model and discussing the trade-offs involved in aligning relevance, scalability, and reliability in a high-stakes, large-scale digital certification environment.

Form Assembly: An AI Agent Workflow for Building Parallel Spring Forms for a Diagnostic Classification Assessment

Ann Hu

Operational test form assembly can be framed as a constrained optimization problem in which assembled forms must satisfy strict blueprint specifications, psychometric quality criteria, and linking requirements to support scale comparability. This study presents an AI agent to automate the assembly of two parallel forms for a diagnostic classification assessment of foundational reading and phonics skills. The assessment blueprint comprises 11 instructional sections (e.g., Letter Sounds, High-Frequency Words, Consonants and Short Vowels, etc.). The agent constructs two spring forms (3A and 3B), each with 88 unique items, while enforcing exact section and subsection counts, within-form uniqueness, and anchor/overlap requirements across fall, winter, and spring administrations.

The AI agent is defined as a rule-based, goal-directed system that executes a formal assembly specification through iterative actions (select–place–validate–repair) and produces fully auditable outputs. The specification encodes operational constraints and preferences, including: (a) fixed test length and item uniqueness; (b) controlled sourcing from prior forms versus a new item bank; (c) coverage requirements for new-bank items; (d) global and section-level overlap thresholds for anchor design; (e) psychometric quality filters (e.g., difficulty and corrected item–total correlation); and (f) positional constraints to preserve intended instructional progression and avoid anchor clustering.

The assembly process follows a reproducible build–check–repair loop. The agent initializes forms using designated new-bank items, applies a fixed-seed partition to establish parallel structure, and completes forms by sampling from prior administrations and remaining new items. Automated validation routines evaluate all constraints, after which the agent performs targeted swaps until a feasible solution is obtained. Outputs include form-level layouts, overlap diagnostics, and summary statistics suitable for technical review.

This work demonstrates a practical framework for formalizing assembly specifications as machine-executable rules, enabling scalable, reproducible, and transparent test construction while preserving psychometric defensibility through structured human review.

A Stochastic Constrained Test Assembly Method Applied to Multistage Adaptive Testing

Alina von Davier

Multistage testing (MST) assembly is typically treated as a static optimization problem, with panels built once from a fixed pool of calibrated items. In AI-enabled assessment environments, this assumption no longer holds: item pools evolve continuously as automatically generated items enter with uncertain psychometric parameters, and panels must be assembled on demand. MST assembly thus becomes a sequential decision-making problem under uncertainty—which panel should be deployed now to maximize measurement precision, satisfy content and routing constraints, maintain pool sustainability, and accelerate calibration of new items? This presentation extends the Stochastic Constrained Hybrid (SCH) framework, originally developed for linear test forms, to multistage testing. SCH recasts assembly as a multi-armed bandit problem with Fisher information as the reward, balancing exploitation of well-calibrated items against exploration of uncertain ones. In the MST setting, modules serve as the unit of assembly, with stage-level difficulty targets and routing structure entering as additional constraints, while the bandit mechanism governs the placement of uncertain items across stages. A simulation study illustrates the framework's performance relative to conventional assembly approaches, including the shadow-testing method. The main contribution is a principled method for incorporating items with uncertain parameters into automated MST panel assembly.

Designing multi-stage CAT assessments of academic English language proficiency to give examinees the greatest opportunity to demonstrate that they meet performance level thresholds

Stephen Walker

At the English language teaching center at the University of Queensland, we deliver academic English language programs for international students who use English as an additional language. These students have already met the academic prerequisite requirements for university degree commencement but have not yet met the English language proficiency level specified for entry into their chosen degree courses. The cohorts are largely homogeneous in ability meaning that many students are clustered around the critical proficiency level set as the 'minimum entry' level for Australian degree courses.

Such a truncated proficiency profile, missing both high and low ability learners, poses multiple challenges for test designers. If most students are clustered just at, above or below the minimum proficiency level required, how can assessments differentiate well between the abilities of learners when the cohort itself shows little variation in ability?

In this presentation we outline the development process followed to meet this challenge. We set out our multi-stage CAT design for assessments of both academic reading and listening proficiency which employ a series of polytomous testlet stages each consisting of a long stimulus text and an accompanying set of 6 items dependent on that text. We demonstrate how we avoid possible distortions such as inflated reliability estimates and inflated latent trait precision measurements due to local item dependence between testlet sub-items.

We then show how our use of a Bayesian Maximum Falsification approach to next testlet-item selection permits us to:

- improve the test taker experience by including testlets where we expect near full success rather than the expected 50% success of a maximum information approach
- deliver a test where testlets are selected to continuously challenge the examinee to falsify our existing measure estimate thereby giving each student the maximum opportunity to demonstrate their true performance level

In addition to these test taker benefits, we show how employing this test design in our context can reduce the item exposure that would occur under the statistically optimal maximum information approach in which always selecting the "best" item can lead to test-takers of similar ability responding to the same set of targeted items.

References:

Linacre, J.M. Computer-Adaptive Testing, CAT: A Bayesian Approach: Rasch Measurement Transactions, 1995, 9:1 p.412

Lange R. Binary items and beyond: a simulation of computer adaptive testing using the Rasch partial credit model. Journal of Applied Measurement. 2008;9(1):81-104.

parATA: an R-package for Automated Test Assembly

Angela Verschoor

Although Automated Test Assembly (ATA) is a development that started already several decades ago (Theunissen, 1985), the majority of published articles describe optimization models for single test forms, or for a relatively small number of parallel test forms without overlap.

Yet, in our experience, most real-life problems faced by testing agencies involve large sets of parallel test forms from an item pool that is not large enough so that overlap must be allowed. Most tools available for ATA are based on MIP-solvers, which need to introduce large numbers of dummy variables and constraints in the case of large-scale test assembly allowing overlap. To give an example: the assembly of 50 test forms from an item pool of size 1500 while allowing overlap would introduce 1,837,500 binary dummy variables and 1,838,725 related constraints, causing serious challenges even for the best commercial solvers.

Using local search algorithms, for example based on simulated annealing or genetic algorithms, is a viable answer but rarely seen in available tools. parATA is an R-package based on simulated annealing that is currently used for the assembly of up to 60 test forms from item pools containing up to 4000 items.

Compared to tools based on commercial MIP-solvers, parATA is rather slow for the assembly of single test forms, or small numbers of parallel forms without overlap. Typically, MIP-solvers show response times for assembly of, say 5, parallel forms without overlap well below 0.25 seconds, while response times with the parATA package are up to approximately 10 seconds, yielding identical or comparable solutions. On the other hand, assembly of 50 parallel test forms out of a pool comprising 1500 items will typically need 5-10 minutes response time for the parATA package, while optimization using Gurobi was interrupted after 96 hours without yielding a feasible solution at all.

Currently, parATA supports various models for calibration designs, for using Classical Test Theory or Item Response Theory, including a model for the assembly of Multistage Tests. The implemented algorithm seems to be limited to the simultaneous assembly of 60-100 parallel test forms. When more parallel test forms are needed – one of the early adopting clients of parATA assembles 500 test forms per year, and has the ambition to upscale this number to 5,000-10,000 test forms – a Linear-on-the-Fly (LOFT) approach can be considered without compromising reliability or the stability of pass rates, while observing a slightly reduced control over the test overlap.

Beyond Correct and Incorrect: A Nested Logit Computerized Adaptive Test Using Distractor Information to Assess Reading Proficiency

Thiago Costa (co-author: Ocimar Munhoz Alavarse)

This study is based on the development of TAI-Leitura, a computerized adaptive test (CAT) designed to estimate reading proficiency while using information from every response option in multiple-choice items. Conventional educational CATs typically score responses dichotomously, treating all distractors as equivalent. The proposed approach instead models distractor choices as evidence about students' partial knowledge and uses that evidence in proficiency estimation, item selection, and pedagogical interpretation.

The item bank was constructed from Provinha Brasil items, a large scale reading test disposed by the Brazilian Ministry of Education from 2008 to 2016, and its responses data collected in the São Paulo municipal school system between 2012 and 2015, supplemented by item pretest parameters. Rasch, two-parameter logistic, nominal response, and nested logit models were compared in R using the mirt package. The nested logit model was selected because it preserves a logistic component for correctness while modeling the distractors nominally. A bank of 120 items was placed on a common scale using a simulated reference sample of 100,000 response vectors. Two adaptive algorithms were then implemented. TAI-Alt selected the next item by maximizing Fisher information at the current proficiency estimate. TAI-pseudoB additionally used an option-specific "difficulty" index, defined by the intersection between each distractor response curve and the correct-option curve, to guide item selection. Both algorithms used mixed stopping rules: 8 to 20 administered items, standard error below 35 points, and a confidence interval contained within one of the five proficiency levels scale.

Simulation studies and school-based pretests preceded a field administration involving 2,052 students in Grades 1-9 from three public schools. The polytomous models yielded more test information than the Rasch and 2PL models, while the nested logit and nominal models produced similar information. The simulated linking procedure reproduced the original item difficulties almost perfectly ($r = .999967$). In the field administration, participation reached 90.5%, students were classified across all five proficiency levels, and tests administered 14.8 items on average. Mean test lengths by proficiency level were 17.4, 15.4, 19.4, 20.0, and 11.9 items, respectively. Distractor-sensitive response trajectories differentiated students within the same proficiency level and supported individualized interpretations of what students could and could not yet do in reading.

The findings provide a feasibility demonstration for incorporating distractor information into educational CAT. They also reveal important constraints: the item bank offered insufficient information at its upper end, Level 4 students frequently reached the 20-item limit, and the pseudoB algorithm was administered to only 50 students in the main study. Expanding the bank and conducting controlled comparisons of the two algorithms are therefore priorities for further validation.

Computerized Adaptive Testing Using Nonparametric Item Response Theory Models

Cecilia Marconi

Computerized adaptive test (CAT) has consolidated as an efficient tool for measuring abilities. However, its development has historically been linked to parametric item response theory (IRT) models. Therefore, this dissertation explores an alternative approach based on nonparametric models, particularly kernel-based nonparametric regression and the nonparametric isotonic model, which imposes a monotonicity constraint on item characteristic curves (ICCs). The study was conducted under three scenarios: (i) establishing the correct specification between the data-generating (logistic) model and the model used to estimate the ICCs; (ii) including the misspecification between the logistic data-generating model and the logistic model used for ICC estimation; and (iii) determining the presence of logistic curves that deviate from traditional parametric forms. Within this framework, a central challenge lies in the fact that ICCs do not assume predefined functional forms, which limits the use of classical item selection criteria based on Fisher information. To address this issue, item selection procedures based on global information measures were implemented, particularly Shannon entropy and Kullback–Leibler divergence. In addition, pseudo-information was employed as a functional analogue of Fisher information in the context of the isotonic model. Simulation studies are conducted to evaluate the performance of nonparametric approaches in comparison with parametric ones, along with an application of the isotonic model in a real educational setting. The results show that, although parametric models achieve the highest accuracy under correct specification, nonparametric approaches exhibit comparable performance. Under conditions of model misspecification or non-standard ICCs, the nonparametric approaches demonstrate greater accuracy and stability in ability estimation. Furthermore, item selection criteria based on global information measures outperform those based on local information. As a central contribution, a nonparametric CAT based on the isotonic model is implemented within a new technological architecture of Uruguay's national educational assessment system, with concrete applications in real educational contexts. The results provide evidence of precise ability estimates, stability in adaptive test administration, adequate levels of convergent validity in comparison to a previous version of the test, response times compatible with operational use, and a satisfactory testing experience for examinees. Overall, the findings indicate that nonparametric models constitute a methodologically robust and operationally viable alternative for the development of CAT.

Addressing Misclassification Bias in Computerized Adaptive Testing with Latent Performance Standards

Bartosz Kondratak

Many popular approaches to standard setting such as Angoff-based methods, or the Bookmark method (Item Mapping), are designed to establish cut-off scores directly on the latent scale of ability. However, it is widely acknowledged that applying IRT to score test-taker performance relative to such latent cut-offs introduces a risk of misclassifying test-takers due to the statistical bias of the ability estimators. While the statistical properties and distinct bias patterns of various IRT estimators; such as maximum likelihood (ML), Warm's weighted likelihood (WLE), maximum a posteriori (MAP), and expected a posteriori (EAP); have been thoroughly investigated over the years, particularly in the context of CAT, research explicitly quantifying the resultant misclassification effects or proposing methodological fixes remains scarce.

In this presentation, we report the results of a simulation study explicitly designed to investigate how the bias of IRT ability estimators translates into bias in test-taker classification or pass/fail decisions when the cut-off scores are defined on the latent scale. Three estimators are considered (ML, EAP, and Warm's WLE), and we manipulate the average discrimination of items in the pool, the size of the SE stopping criterion, and the placement of the cut-offs. A straightforward Monte Carlo procedure to reduce this misclassification bias is also proposed and evaluated. We first transform the latent cutoff θ_c onto the expected scale of the IRT ability estimator, $\theta^*c = E(\theta^\wedge|\theta_c)$, and then use that transformed θ^*c to make decisions about $\theta^\wedge$. This first-order correction for estimator bias around θ_c is shown to be an effective technique for the removal of classification bias in the analysed settings, specifically for EAP, where this problem was most prominent.

We observed that the transformation $E(\theta^\wedge|\theta_c)$ did not depend on the average discrimination of items in the pool; however, it was impacted by the value of the SE stopping criterion. This is in accordance with theoretical expectations for EAP because its bias is driven by the ratio of the test information to the variance of the prior distribution and the fixed-SE stopping rule forces the former to be constant. The practical consequence is that if a given CAT is being used with a varying SE stopping criterion, we should either apply transformed cutoffs that are specific to each stopping rule value or resort to ability estimators that suffer the least from bias.

Testlet-Based Termination in Relaxed CAT: Reducing Assessment Length and Testlet Exposure

Luciana Cancado (co-authors: Kristin Morrison, Yu Su)

Operational interim assessments administered across multiple testing windows impose competing demands that standard computer adaptive test (CAT) item selection handles poorly: controlling item repeat rates across a single item bank, aligning content to student on- and off-grade level, and balancing measurement precision with administration burden.

Cappaert, Morrison, and Su (2022) introduced Relaxed CAT (R-CAT), a modified Rasch-based b-matching approach with three components: standard b-matching difference, a recency-weighted repeat penalty, and a grade-level preference penalty. Simulations based on an operational interim assessment demonstrated substantial reductions in item repeat rates, improved student-to-item grade alignment, and maintained reliability and conditional standard error of measurement comparable to unconstrained b-matching. However, R-CAT did not address test length or duration.

This proposal extends R-CAT by integrating a strategy for shortening the assessment that includes (1) reducing items administered per domain and (2) variable-length domain stopping rules. The latter uses two types of domain-specific stopping rules. Domains with typical item-based selection use a standard error of measurement (SEM)-based termination logic. For passage-dependent reading comprehension domains with testlets, we introduce a novel testlet-based termination logic. Specifically, a testlet-based domain terminates when either (a) the minimum domain item count is reached and all items in the most recent passage exhausted, or (b) a maximum domain item count is reached, even if only a subset of items from a testlet was administered. This logic addresses an operational constraint: students should not begin a new testlet when information gain is minimal. This study evaluates how the proposed testlet-based termination logic helps reduce test length beyond a straightforward reduction in item count for testlet-based domains within the R-CAT framework through simulation.

To evaluate the effect of the testlet-based termination logic on test length, we compared median length of testlet-based domains with the maximum number of items that would have been administered in a fixed length format and percentage of students seeing less than the maximum number of items for testlet-based domains. We also looked at passage exposure of shortened R-CAT compared to a baseline condition that only uses unconstrained b-matching for item selection (i.e., full length test without penalty weights). Measurement precision is evaluated through marginal reliability of both overall reading test and reading comprehension domain.

Simulation results indicate that the testlet-based termination logic further reduced the median length of testlet-based domains by 2 items in grade 4, from a maximum of 24 to 22 items for the shortened test, and about 75% of students saw less than the maximum number of items. The shortened R-CAT also significantly reduced number of reading passages administered when compared to baseline, e.g., over 40% of grade 4 students received six passages compared to a wider distribution at baseline (see

REF_Ref230964926\h\MERGEFORMATFigure108D0C9EA79F9BACE118C8200AA004BA90B02000000080000000E0000005F0052006500660032003300300039003600340039003200360000000). Both overall and domain marginal reliability remained high (over .94 and .91 respectively), only marginally below baseline (see REF_Ref230869278\h\MERGEFORMATTable108D0C9EA79F9BACE118C8200AA004

BA90B02000000080000000E0000005F005200650066003200330030003800360039003200370038000000).

These results indicate that incorporating testlet-based domain stopping rules to R-CAT helped further reduce average domain length while preserving measurement precision. Additionally, testlet exposure is reduced, increasing item bank availability for future administrations.

Computer Adaptive Testing with small item bank: Improved precision and reduced opportunities for cheating

Rense Lange (co-authors: Haniza Yon, Labiib Marzuki, Amrit Lakra)

Computer Adaptive Testing (CAT) is based on the principle that test items are selected from an item bank to maximize the information they provide about an examinee's ability or latent trait. Although the literature generally recommends that item banks should contain at least 100 items for effective CAT, such large item banks are rarely available when developing new CAT applications. In this study, we investigated whether CAT remains beneficial with much smaller item banks. We tested two hypotheses. First, we hypothesized that CAT would yield more precise ability estimates even when only 15–30 items are available in the item bank. Second, we hypothesized that, even if the gain in precision is modest, CAT could reduce opportunities for cheating in group testing by presenting different item sequences to neighboring examinees. These hypotheses were evaluated using two datasets: (A) a 30-item cognitive ability test and (B) a 16-item psychiatric assessment. Results supported both hypotheses. In both datasets, maximum-information item selection consistently yielded more precise ability estimates than did the random selection of items. To evaluate test security, we repeatedly sampled pairs of examinees and compared the order of item administration using Spearman's rank correlation coefficient (ρ). For both random item selection and maximum-information CAT, the average ρ between the ordering of the CAT administered items was approximately zero. Thus, neighboring examinees will receive essentially unrelated item sequences, making answer copying substantially more difficult. We conclude that CAT can be useful even with very small item banks, both by improving measurement precision and by reducing opportunities for cheating.

Posters

Applications and Implications of Unidimensional, Multidimensional, and Bifactor Models in Measuring Students' Socioeconomic Status

Joyce Soares Rodrigues Petrus (co-author: Flavia Pereira Xavier)

Socioeconomic status (SES) index are widely used in large-scale educational assessments to represent social inequalities associated with academic performance. These indicators are typically constructed from self-reported items concerning household assets, access to services, housing conditions, and parents' educational attainment, and are often treated as measures of a single latent trait. In Brazil, these data are collected through the contextual questionnaire of the National Basic Education Assessment System (Saeb), administered to elementary and higher education students.

SES estimation using Saeb data is usually based on Item Response Theory (IRT) models that assume unidimensionality. However, the content of the questionnaire items suggests the presence of distinct dimensions of family background, such as economic resources, parental education, and housing conditions. This issue is particularly relevant because different model specifications may affect score estimation, student classification, and the interpretation of the relationship between SES and academic achievement.

This study compares unidimensional, correlated multidimensional, and bifactor models to investigate whether the multidimensionality of SES is substantively meaningful or whether a general factor is sufficient to represent the social inequalities associated with educational outcomes.

To address this question, we use the 2023 Saeb database, the most recent publicly available dataset, comprising slightly more than 7 million students, of whom 5.7 million answered the contextual questionnaire. Unidimensional, correlated multidimensional, and bifactor graded response models were estimated to evaluate the adequacy of the unidimensionality assumption, the relevance of specific dimensions of family background, and the existence of a general SES factor. All analyses were conducted using the mirt package in R.

Model comparisons focused on the psychometric adequacy of the estimated structures, the stability of student classifications based on SES scores, and the predictive validity of the resulting scores with respect to academic achievement.

The results suggest that, although multidimensional and bifactor structures provide a more detailed representation of specific components of students' socioeconomic backgrounds, a general SES measure offers an adequate representation for broad educational analyses. This study contributes to the literature by discussing the limitations of unidimensional representations of SES and by empirically evaluating the analytical benefits associated with the use of more complex psychometric models in large-scale educational assessments.

Can Large Language Models Predict 3PL Item Parameters in Brazilian Portuguese? An Exploratory Study with ENEM Items

Pedro Otavio Mantovani (co-author: André Carlos Ponce de Leon Ferreira de Carvalho)

Item calibration under Item Response Theory (IRT) traditionally relies on large pilot samples, making the development of educational assessments costly and operationally demanding. While most research on item parameter prediction has focused on item difficulty and relied on traditional natural language processing (NLP) approaches, considerably less attention has been devoted to predicting complete three-parameter logistic (3PL) item characteristics using modern Large Language Models (LLMs). Furthermore, most existing studies have focused on English-language assessments, leaving open questions regarding the applicability of these approaches to other languages and educational contexts.

This exploratory study investigates whether state-of-the-art LLMs can estimate 3PL item parameters directly from the textual content of Brazilian Portuguese assessment items. The study uses items from the Brazilian National High School Exam (ENEM), one of the world's largest university admission exams and a large-scale assessment calibrated under IRT.

A few-shot prompting approach was adopted. Three commercial LLMs (ChatGPT, Claude, and Gemini) received nine previously calibrated ENEM language items as examples, selected to represent diverse psychometric profiles, and were asked to estimate discrimination (a), difficulty (b), and guessing (c) parameters for 35 unseen items from the 2024 exam. The predictive performance obtained was evaluated using Pearson correlation, coefficient of determination (R^2), mean absolute error (MAE), root mean squared error (RMSE), and bias.

The experimental results indicate that all models showed substantially stronger performance for item difficulty than for discrimination or guessing parameters. Difficulty prediction achieved R^2 values ranging from 0.25 to 0.46, with the highest value obtained by Gemini ($R^2 = 0.455$). In contrast, discrimination and guessing estimates exhibited near-zero predictive power and a tendency toward regression to the mean. Despite these limitations, the predictive performance observed for difficulty is broadly consistent with levels reported in previous studies using machine learning and NLP-based approaches.

Although current results remain insufficient for operational item calibration, the findings provide initial evidence that contemporary LLMs can capture psychometrically relevant information from Brazilian Portuguese assessment content. The study contributes to the emerging literature on AI-assisted psychometrics by extending parameter prediction research beyond English-language contexts and by evaluating the feasibility of estimating complete 3PL parameter sets using general-purpose LLMs. Future work will explore the use of larger sets of calibrated items, fine-tuning strategies, in addition to the use of LLM as auxiliary information in psychometric calibration procedures and interpretable predictive models.

CAT API: Exploring the Application of Computerized Adaptive Testing for Formative and Summative Assessments

Bruno Silvestre (co-authors: Thiago Ferreira Miranda, Mariana Cúri)

This work presents the development and evaluation of a computational solution for applying Computerized Adaptive Testing (CAT) in the context of formative and summative assessments. The proposed solution consists of an open-source educational API that integrates Item Response Theory (IRT) models and Cognitive Diagnostic Models (CDM), well-established methodologies in the field of Psychometrics. The study addresses both theoretical and computational aspects, including the investigation of psychometric models, item exposure and content balancing criteria, as well as the implementation of a modular architecture using the Django framework in Python and R packages specialized in IRT and CDM.

The solution was designed to reduce technical and conceptual barriers related to the adoption of adaptive assessments by providing low integration complexity and support for both proficiency estimation and the diagnosis of specific student skills. The system also includes item bank management mechanisms and an administrative interface for educational applications.

The evaluation comprised simulation-based experiments and practical classroom applications, enabling the analysis of the tool's behavior in real educational scenarios. The results indicate that the proposed approach is feasible, flexible, and capable of providing relevant and more accurate information to evaluators, contributing to the broader adoption of adaptive assessments in different educational contexts.

Comparing Large Language Models for Psychological Scale Translation: Translation Quality Scores and Blinded Item Selection in the Brazilian Adaptation of the Teacher Autonomous Behavior Scale

Thales Araújo Dias (co-author: Francisco Pablo Huascar Aragão Pinheiro)

The International Test Commission (ITC) Guidelines for Translating and Adapting Tests recommend that translators possess knowledge of the languages, culture, test content, and assessment principles. In practice, this combination is difficult to meet and researchers sometimes resort to translating from a third language rather than directly from the source, adding a layer of distortion. Large Language Models (LLMs) have been proposed as an alternative, given their ability to incorporate construct definitions and methodological guidelines directly in the prompt. The Teacher Autonomous Behavior Scale (TABS) is a case in point: originally developed in Dutch, a language for which finding translators fluent in Brazilian Portuguese with psychometric expertise is challenging. This study used the TABS adaptation to compare LLM-generated translations and investigate how automated quality metrics relate to human judgments in item selection. Method: Fifteen models from five LLM families were used: Claude (Haiku 4.5, Sonnet 4.6, Opus 4.6), Gemini (Flash Lite 3.1, Pro 3.1), Gemma (2:9B, 3:4B, 4: E4B, TranslateGemma:4B), ChatGPT (5.4, 5.4 Mini, 5.4 Nano, 5.4 Pro), and Ollama (3.1:8B, 3.2:3B), the most advanced versions available in March-April 2026. All received identical prompts with ITC guidelines, construct definition, and psychometric background, set to temperature 0.0, except GPT 5.4 Pro. LLM calls were made via the *ellmer* package (v0.4.1) in R; translation quality was assessed using the Crosslingual Optimized Metric for Evaluation of Translation (COMET) in reference-free mode (*wmt22-cometkiwi-da*) via Python (v3.11.9) with the *comet* package (v2.2.7). Differences were examined via Generalized Estimating Equations (GEE); Gaussian structure showed better fit than Gamma (QIC=30.2 vs. 555.0), followed by pairwise Tukey comparisons. Models without significant pairwise differences were retained for blinded item selection, in which authors chose the most appropriate version without knowledge of the source model or COMET score, based on semantic, idiomatic, experiential, and conceptual equivalence criteria. The final list was reviewed by a native Dutch speaker fluent in Brazilian Portuguese, followed by evaluation from three psychological assessment experts. Results: All 15 models yielded mean COMET scores in the good range (0.80-0.94). The highest means were GPT 5.4 Nano and Claude Haiku 4.5 (both $M = 0.890$); Ollama 3.2:3B produced the lowest ($M = 0.862$). GEE indicated a significant overall effect (Wald = 64.2, $p < .001$), but no pairwise differences survived Tukey correction ($p \geq .271$), so all models were retained. The final 20-item version achieved a mean COMET score of 0.881 (SD=0.023), with items drawn from 13 of the 15 models; GPT 5.4 Pro contributed the most ($n = 7$), followed by Claude Sonnet 4.6 ($n = 3$). The Dutch reviewer approved all translations without modifications. Three experts suggested adjustments, including implicit subject phrasing in item stems. Discussion: Multiple LLMs generated a wider candidate pool per item than traditional designs. Higher-scoring models were not necessarily preferred in blinded selection, suggesting COMET scores alone are insufficient for item choice. Expert adjustments underscore that human judgment remains essential. Still, LLMs may serve as a useful starting point for adapting instruments originally developed in less widely spoken languages.

Developing a German reading comprehension CAT for learning progress assessment

Ophelia Urbach (co-authors: Rebecca Schneider, Elmar Souvignier)

As reading competence in German elementary students overall declined while heterogeneity increased (Frey et al., 2023; Schneider & Wittig, 2022), teachers are challenged to fit instruction to the childrens' needs. Formative assessment (e.g., using the approach of Learning progress assessment (LPA)) is a promising way to provide individually tailored support measures based on diagnostic information. For regular implementation in the classroom, the single tests need to be short and easy to administer by teachers, while also meeting psychometric criteria (Souvignier et al., 2021). Given the increasing heterogeneity of student performance, conventional methods (i.e. parallel testing) are reaching their limits: Achievement tests must include a sufficient number of items at all levels of difficulty to account for the variation in student performance (Moosbrugger & Brandt, 2020). In this regard, computerized adaptive testing (CAT) provides an opportunity to achieve high measurement efficiency (Frey, 2012), while simultaneously covering a broad range of abilities and thus address the high degree of heterogeneity. The aim of our study is to develop and evaluate reading comprehension tasks suitable for CAT as well as for LPA for third and fourth graders. Our research questions are related to a newly developed item pool of about 300 self-developed items. Each item consists of short texts (50-150 words) with one or two corresponding reading comprehension tasks. First, we wanted to analyse if a one-dimensional model has a sufficient model fit. Second, we investigated if the difficulty range of the item pool is sufficient for adaptive testing. To pilot the reading comprehension tasks, we drew upon the design principles of large-scale assessments and used a balanced, incomplete multiple-matrix sampling design (Gonzalez & Rutkowski, 2010). We created 25 Item blocks which we distributed in a rotating order across 25 paper-and-pencil booklets, each containing 4 blocks. Testing began in October 2024. Due to systematic missing values at the end of each block in the first wave ($n = 650$), a second wave is currently being conducted with the order of items reversed within the blocks. Initial IRT- analyses of the first wave show a better model fit for the one-dimensional 1PL-Model than the 2PL-Model ($\chi^2(252) = 824.77$, $p < .001$) and the 2PL-Model better than the 3PL-Model ($\chi^2(253) = 326.89$, $p = .001$). The range of item difficulties varied between $-2.99 < \theta < 2.35$. Encouragingly, the preliminary results indicate a broad range of item difficulties. Furthermore, they support a one-dimensional model. However, results of the combined waves of testing are still pending, as is the analysis regarding students' grades, missing values and items with conditional responses. As a mid-term perspective, the items shall be implemented in the digital LPA-system and will be evaluated under real classroom conditions.

Psychometric Calibration of the Adolescent Externalizing Behaviors Bank (AEBank) Using the Graded Response Model: Item Functioning, Information, and Differential Item Functioning by Sex

Sebastian Jesus Garrido (co-authors: Fernanda Belén Ghio, Ana Estefanía Azpilicueta, Valeria Morán, Manuel Bruzzone, Marcos Cupani)

Externalizing behaviors during adolescence represent a major public health concern due to their association with violence, substance use, school dropout, and antisocial trajectories. Accurate assessment of these behaviors requires measures capable of providing precise information across varying levels of severity. Item Response Theory, particularly the Graded Response Model (GRM), offers a framework for evaluating discrimination, thresholds, and measurement precision across the latent trait. This study calibrated the Adolescent Externalizing Behaviors Bank (AEBank; 53 items) in a sample of 980 adolescents from Córdoba, Argentina. Analyses were conducted using the GRM implemented in the mirt package in Rstudio (2026.5.0.218). Global fit indices showed mixed evidence: $M^2 = 5706.31$ ($df = 1325$, $p < .001$), RMSEA = .058 (90% CI [.057, .060]), CFI = .911, TLI = .908, SRMSR = .098. While RMSEA and incremental indices suggested acceptable scale-level fit, SRMSR exceeded conventional thresholds ($< .08$), and several Q3 residual pairs (e.g., items 8–9, 15–16, 17–19) exceeded .20, indicating localized dependence, potentially reflecting shared item content. Item fit was evaluated using the S- χ^2 statistic. Most items demonstrated adequate fit. Items 17, 19, 44, and 46 showed statistically significant misfit, though associated RMSEA values (.037–.044) suggested limited practical impact. Discrimination parameters ranged from moderate to very high ($a = 0.586$ – 5.279), with several items exceeding > 2.0 . Threshold parameters (b) spanned a wide range of the latent continuum ($b_1: -1.71$ to 3.41 ; $b_4: 0.79$ to 7.97), with denser coverage in the moderate-to-high severity region ($\theta \approx 1$ to 3), consistent with the Test Information Function, which peaked around $\theta \approx 2$. Together, these findings support efficient item selection for Computerized Adaptive Testing (CAT) targeting at-risk adolescents. Marginal reliability was high ($rx = .891$), with comparable empirical reliability based on EAP scores ($rx = .896$). Differential item functioning (DIF) by sex was assessed using ordinal logistic regression via the lordif package, employing χ^2 tests ($\alpha = .01$) and McFadden pseudo- R^2 effect-size criteria ($\Delta R^2 \geq .02$). Fifteen items were excluded due to sparse response frequencies. Although 19 of 38 items were statistically significant, most effects were negligible. Monte Carlo simulations (1,000 replications) suggested that many flagged effects fell within empirical thresholds expected under the null hypothesis. A limited subset (items 1, 19, 22, 26, 35, 37) showed comparatively larger effects. Item 26 exhibited the strongest uniform DIF ($\Delta R^2 = .129$), indicating systematic differences in item endorsement across sexes at equivalent trait levels. Overall, findings support the AEBank as a psychometrically adequate item bank for assessing externalizing behaviors in adolescents. Its optimal precision at elevated trait levels supports its use in screening contexts and future CAT applications targeting adolescents at higher behavioral risk. Although minor DIF by sex was detected, practical impact was limited, supporting broad score comparability. Items 1, 22, and 26 should be prioritized for review in future refinements. Subsequent research should explore testlet or bifactor representations to account for local dependence and extend DIF analyses to additional sociodemographic groups.

Psychometric Calibration of the CIMEC-40 for Eating Disorder Risk Assessment in Adults: Development of a Screening Tool

Ana Estefanía Azpilicueta (co-authors: Fernanda Belén Ghio, Sebastian Jesus Garrido, Valeria Morán, Marcos Cupani)

Among adults, eating disorders represent a significant mental health concern. Early detection and intervention require brief, psychometrically sound screening instruments capable of differentiating individuals according to their level of risk. One upstream factor is sociocultural pressure toward thinness, which precedes body dissatisfaction and eating pathology. This study aimed to calibrate the CIMEC-40 as the first screening module of a Computerized Adaptive Testing (CAT) battery designed to assess eating disorder risk. The sample included 489 Argentine adults ($M = 22.63$ years; 76.3% female). Item calibration was performed under the Graded Response Model (Rstudio 2026.04). Model fit was evaluated using M_2 statistic, CFI, TLI, RMSEA, and SRMR. We examined item discrimination (a), thresholds (b), test information function, reliability (EAP rxx), local dependence (Q3 residuals $> .20$). Differential item functioning (DIF) by sex was examined using logistic regression procedures (χ^2 test, $\alpha = .01$). DIF magnitude was evaluated using changes in pseudo- R^2 values, with $\Delta R^2 < .035$ indicating negligible and $> .070$ large DIF. Global fit indices supported a unidimensional model: $M_2 = 2013.53$ ($df = 620$, $p < .001$), CFI = .913, TLI = .907, RMSEA = .068 (90% CI [.065, .071]), SRMR = .074. The mean Q3 correlation was -.012, indicating no overall local dependence. However, two redundant item pairs within the Advertising Influence factor (items 6–8: 0.625; 23–26: 0.503) showed notable dependence, which may be due to content overlap rather than multidimensional structure. Discrimination parameters (a) ranged from 0.510 to 2.662. The strongest items were 8 (2.662), 28 (2.572), 11 (2.550), and 6 (2.518), mostly from the Advertising Influence factor. The weakest items were 39 (0.510), 24 (0.604), and 27 (0.681). All thresholds were found to maintain the expected increasing order, confirming the proper ordering of the response categories. The observed ranges (b_1 : -3.36 to 1.10; b_4 : 1.17 to 5.79) indicate that the items cover a broad spectrum of the latent trait, from low to very high levels. The test information function indicated maximum precision at moderate-to-high trait levels ($\theta \approx 1-3$), and reliability was high (EAP rxx = .955). Seven items were flagged for DIF by sex (items 5, 16, 17, 21, 27, 29 and 39). However, pseudo- R^2 change was moderate for these items and large only for Item 5. Based on these results, future ROC analyses could establish optimal cut-off scores for screening. Participants below the cut-off would receive psychoeducational feedback; those above would proceed to body dissatisfaction and neuroticism assessment modules, enabling stepped care according to risk level. The evidence suggests that the CIMEC-40 can be used operationally as a largely unidimensional scale for assessing sociocultural pressure toward thinness in adults. Its adequate precision in moderate-to-high trait ranges and high reliability support its use as an efficient first-step screening tool for adults at risk for eating disorders. Items with lower discrimination parameters and large DIF should be prioritized for revision, while redundant item pairs should be carefully evaluated in future adaptive testing designs. Findings support future CAT development.

Psychometric Evaluation of a Social Competence Item Bank for Children: A Rasch Partial Credit Model and CAT Simulation Approach

Valeria Morán (co-authors: Manuel Bruzzone, Sebastian Garrido, Fernanda Ghio, Ana Azpilicueta, Marcos Cupani)

Social competence, defined as the ability to achieve personal goals in social interactions while maintaining positive relationships with others (Gresham, 2001), represents a central domain of child development with direct implications for school adjustment, mental health, and academic performance. Despite its relevance, rigorous psychometric evaluation of social competence in Spanish-speaking children using Item Response Theory (IRT) remains limited. Most validation studies in Latin American populations have relied exclusively on confirmatory factor analysis and internal consistency indices, without leveraging the measurement invariance properties that IRT models afford. This study presents the first systematic IRT-based calibration of a 53-item bank designed to assess social competence in school-age children, using the Partial Credit Rasch Model (PCM) and a post-hoc Computerized Adaptive Testing (CAT) simulation. A sample of 1,214 children (51.2% male; mean age = 9.8 years, SD = 1.4; range 7–12) from public and private primary schools in Córdoba, Argentina, completed a self-report measure operationalized through behaviorally-specific items across six facets of social competence: Assertiveness, Conflict Resolution, Self-regulation, Conversation, Civility, and Empathy. Analyses were conducted using RUMM2030, following established Rasch calibration guidelines. Results supported the overall psychometric adequacy of the item bank. Unidimensionality was confirmed via Principal Component Analysis of Residuals (3.72% significant t-tests, below the 5% threshold), and no item pairs exceeded the local independence criterion (residual correlations $\leq .20$). The Person Separation Index (PSI = .786) met the recommended threshold for group-level use ($\geq .70$), indicating that the bank adequately differentiates among children with varying levels of social competence. Item difficulty hierarchy revealed a theoretically coherent gradient: Empathy and Civility items clustered at the easier end of the continuum, while Assertiveness items, particularly self-affirmation behaviors, were the most challenging, consistent with developmental models of social competence acquisition. The CAT simulation demonstrated strong efficiency gains. Under a stopping criterion of $SE \leq .30$ logits, the adaptive algorithm required an average of 11.8 items to achieve precision comparable to the full bank ($r = .962$ with full-bank estimates), representing a 77.7% reduction in test length. A more permissive criterion ($SE \leq .40$) yielded adequate precision with only 7.3 items on average ($r = .941$), suitable for rapid screening applications. These findings support the viability of deploying an adaptive social competence assessment for children in educational settings where administration time is constrained. Future directions include expanding item coverage at the extremes of the latent continuum, incorporating multi-informant versions (parent and teacher report), examining criterion validity against indicators of school adjustment and well-being, and conducting systematic differential item functioning (DIF) analyses across sex, age, and school type to ensure measurement equity across subgroups.

The effect of (mis)judged provisional item difficulties on calibrations: Evidence from online pretesting in Linguaskill and English Skills Test

Dario Delelis (co-authors: Tom Abraham, Tom Iveson-Malcom, Hannah Haynes, John Begley)

In the Rasch model, item difficulty is estimated from a probabilistic logistic function of the difference between candidate's ability (θ) and item's difficulty (δ), but assumes excellent model fit (Benton, 2021) and balanced sampling of the ability distribution in the calibration cohort. In operational practice, new and uncalibrated items may be assigned inaccurate provisional difficulties based on human judgement resulting in exposure to candidates with significantly mismatched abilities. This mistargeting can affect the quality of calibration accuracy (He et al., 2020), which may introduce error in items later used to calibrate others (Fink et al., 2025).

There is little operational evidence regarding the impact of highly misjudged provisional difficulty values on difficulty estimation in live adaptive testing. This work asks whether evidence exists of these aberrant provisional values persisting into final calibrated difficulties under the Rasch model.

In Linguaskill and EST, new items are assigned a provisional difficulty per expert judgement and seeded into live sessions in an online calibration model. Provisional difficulties act as constraints on the candidate range exposed to a given item (Kingsbury, 2009). This affects the composition of the cohort used to estimate the item's difficulty parameter under IRT (e.g. de Ayala, 2021). If provisional item δ is misjudged, this may affect their final calibrated values unless the Rasch model can sufficiently correct for this. To investigate, a set of Linguaskill and EST Listening and Reading MCQ items of known difficulties derived from whole-bank calibration was selected. Items were seeded into live pretesting slots in both tests using new provisional difficulties at +1.5 logits higher (H-version), -1.5 logits lower (L-version), and at their original (control) value (C-version). Items were exposed to a minimum of 250 candidates and calibrated under the Rasch model using the mirt R package (Chalmers, 2012) and MML estimation.

Our findings suggest the presence of a significant but non-uniform compression effect. Across both skills, mean logit δ shifts were negligible for C items, but almost completely undid the +/- 1.5 logit aberration in a large subset of H and L items, compressing them into their original difficulty space. This compression was not strongly moderated by sample size, facility, infit, standard error, or mean candidate θ , and not significantly different between Linguaskill and EST. Absolute δ differences to original values were relatively small on average and infit and outfit were acceptable. However, difficulties remained aberrant for 23% of items, average RMSE was higher than expected, and SEM was significantly higher for both H and L (Listening) and H items (Reading) compared to C.

Results indicate that the Rasch compression effect, while significant, does not uniformly correct misjudged difficulties items, leaving a large subset with misaligned δ s and high SEM and RMSE. This may cause cumulative misalignment effects in adaptive item banks if not accounted for. In response, we propose that more accurate provisional δ estimation for items in adaptive tests (e.g. using machine learning and/or regression models) may be necessary to further reduce the risk of misestimation of final calibrated values.

Organizing Institutions



The University of São Paulo (USP) is a public university maintained by the State of São Paulo. It is consistently ranked among the top universities in Latin America. As the leading institution of higher education and research in the country, USP is responsible for educating a significant share of Brazil's Master's and Ph.D. graduates. As a public and tuition-free institution, it plays a central role in promoting accessible, high-quality education. In numbers, USP has more than 90,000 students, around 5,500 faculty members, 12,500 technical and administrative staff, and offers 185 undergraduate and hundreds graduate programs across 8 campuses



SESI (Social Service of Industry) is a private, nonprofit organization that operates throughout Brazil through a nationwide network of regional units, providing support to industry and its workforce. It plays a fundamental role in Brazil's social development, contributing to the well-being and quality of life of industrial workers, their families, and communities through services in education, health, leisure and sports, culture, food, and more.



The International Association for Computerized Adaptive Testing (IACAT) provides resources and a biannual conference to the community of researchers and practitioners that are driving the next generation of assessment.

Supporting Institutions



**ASSOCIAÇÃO BRASILEIRA
DE ESTATÍSTICA**

Conference Sponsors

Platinum Sponsor



Silver Sponsors



Assessment Systems has been a trusted leader in computerized testing since 1979. We deliver best-in-class software platforms and psychometric consulting services for certification bodies, higher education institutions, government agencies, and employers worldwide. Their scalable solutions encompass item banking, online delivery, adaptive testing, candidate/certification management, remote proctoring, and advanced analytics, all grounded in modern psychometric standards. ASC's mission is simple: better tests lead to better decisions about people, providing positive impact to career and educational opportunities.



**Northwestern
University**

The Division of Outcomes and Measurement Science at Northwestern, based in downtown Chicago, Illinois, strives to improve health/healthcare through state-of-the-science psychometric measurement and evaluation. Our faculty of quantitative/qualitative psychologists, clinicians, and informaticians are pushing the frontiers of health measurement, having created and now distributing multiple tablet and smartphone-based neuropsychological and self-report assessments used in more than 1,100 institutions around the world: NIH Toolbox, Mobile Toolbox, PROMIS, NeuroQoL, MyCog, MyCog Mobile, Baby Toolbox, and more.

Bronze Sponsors



The National Registry of Emergency Medical Technicians (National Registry) is the nation's non-profit, non-governmental EMS certification organization. Dedicated to supporting the EMS profession through partnerships, research, and lifelong assessment of clinical competence, the National Registry helps ensure EMS clinicians are prepared to provide safe and effective patient care throughout their careers.



Prepona is a Brazilian company founded in 2000, specializing in digital assessment solutions. A pioneer in Latin America, it has stood out for innovations such as adaptive testing, digital certifications, and secure remote administration. Internationally recognized, it offers a comprehensive portfolio that includes online testing, test center delivery, and mobile-based assessments, with a focus on security, accessibility, and flexibility for assessment programs.

Custom Sponsors



Centro de Políticas Públicas e Avaliação da Educação da Universidade Federal de Juiz de Fora



O Centro de Treinamento e Desenvolvimento (CETREDE) is a non-profit institution affiliated with the Universidade Federal do Ceará (UFC). Founded in 1964, it emerged from partnerships between the UFC, the Banco do Nordeste, and the Brazilian Government, under a technical cooperation project of the Organization of American States (OAS). Headquartered in Fortaleza, it has operated for over six decades in education, professional training, research, consulting, and project management focused on regional development.



Adaptativa Inteligência Educacional is a Brazilian education technology company founded in 2010, specializing in psychometric assessment and the development of technology-based solutions for education. Throughout its trajectory, the company has built solid experience in projects carried out with public education systems, private institutions, and educational organizations, including large-scale assessment initiatives. Its work integrates expertise in education, statistics, and technology to support the entire

assessment cycle—from the design and qualification of assessment instruments to data processing, analysis, and interpretation—generating evidence to inform decision-making, monitor learning outcomes, and improve educational management.

