PROMS2025 Booklet

The 20th Pacific Rim Objective Measurement Symposium

Next Generation Measurement:
When Innovation Meets Objectivity

July 21-23, 2025

SINGAPORE UNIVERSITY OF SOCIAL SCIENCES

TABLE OF CONTENTS

Programme Overview

21 July 2025, Monday

TIME	WORKSHOP TITLE	VENUE
9:00 AM	Registration	SR.C.3.08
9:30 AM	Workshop 1: Adapting Rasch Meta-metre of Growth for Variables in the Social Sciences **Emeritus Professor David Andrich** The University of Western Australia	SR.C.3.08
9:30 AM	Workshop 2: Item Banking and Adaptive Testing **Dr Rassoul Sadeghi** Australian Curriculum, Assessment and Reporting Authority	SR.C.3.09
10:30 AM	Morning Tea and Networking	SR.C.3.10
12:30 PM	Lunch and Networking	SR.C.3.10 / SR.C.3.11
1:30 PM	Workshop 3: Rasch Measurement for Rater-mediated Assessment **Professor Jue Wang** University of Science and Technology China	SR.C.3.08
1:30 PM	Workshop 4: Rasch-GZ: Item Analysis and Test Equating in the Age of AI **Professor Quan Zhang** Jiaxing University, Zhejiang, China The World Sports University, Macau SAR, China	SR.C.3.09
3:00 PM	Afternoon Tea and Networking	SR.C.3.10
4:30 PM	End of Pre-conference Workshops	

Programme Overview

22 July 2025, Tuesday

TIME	SESSION TITLE	VENUE
8:30 AM	Registration	SR.C.3.11 / SR.C.3.10
9:30 AM	Welcome Address **Associate Professor Allan Chia** Deputy Provost Singapore University of Social Sciences	SR.C.3.08 / SR.C.3.09
9:40 AM	Group Photo, Morning Tea and Networking	SR.C.3.11
10:00 AM	Presidential Address: Measurement, Explanation, and Invariance: Next Generation Invariant Measurement **Professor George Engelhard Jr.** President, Pacific Rim Objective Measurement Society University of Georgia	SR.C.3.08 / SR.C.3.09
11:00 AM	Day 1 Concurrent Session I	Refer to C-01
12:00 PM	Lunch and Networking	C.5.04
1:00 PM	Day 1 Concurrent Session II	Refer to C-02
2:25 PM	Keynote Presentation: Addressing Human Scoring in Subjective Creativity Assessments **Professor Jue Wang** University of Science and Technology China	SR.C.3.08 / SR.C.3.09
3:15 PM	Afternoon Tea and Networking	SR.C.3.11
3:45 PM	Plenary Roundtable Discussion: The Road Ahead: Future Challenges and Opportunities in Objective Measurement **Professor George Engelhard Jr.** University of Georgia **Professor Zi Yan** The Education University of Hong Kong **Professor Quan Zhang** Jiaxing University, Zhejiang, China; The World Sports University, Macau SAR, China **Professor Guanzhong Luo** South China Normal University; Jiangxi Normal University of China **Associate Professor Jeffrey Durand** Toyo Gakuen University **Dr Mohd Zali Mohd Nor** Newstar Agencies Sdn. Bhd. **Dr Iris Lee (Moderator)** Singapore Ministry of Education	SR.C.3.08 / SR.C.3.09
4:55 PM	End of Day 1	
6:30 PM	Conference Dinner	SPGG

Programme Overview

23 July 2025, Wednesday

TIME	SESSION TITLE	VENUE
9:00 AM	Distinguished Scholar Presentation: Rasch Meta-metres of Growth in Reading and Mathematics Attainment at Both Population and Individual Levels **Emeritus Professor David Andrich** The University of Western Australia	SR.C.3.08 / SR.C.3.09
9:50 AM	Morning Tea and Networking	SR.C.3.11
10:15 AM	Day 2 Concurrent Session I	Refer to C-03
11:35 AM	Day 2 Concurrent Session II	Refer to C-04
12:35 PM	Lunch and Networking	C.5.04
1:35 PM	Day 2 Concurrent Session III	Refer to C-05
2:40 PM	Keynote Presentation: What Can Generative AI Do for Assessing Listening and Speaking Skills **Associate Professor Vahid Aryadoust** National Institute of Education, Nanyang Technological University, Singapore	SR.C.3.08 / SR.C.3.09
3:30 PM	Afternoon Tea and Networking	SR.C.3.11
3:55 PM	Keynote Presentation: Refocusing Educational Measurement: Understanding Student Learning through Adaptive Learning **Dr Che Yee Lye** Singapore University of Social Sciences	SR.C.3.08 / SR.C.3.09
4:45 PM	Closing Remarks **Professor Zi Yan** Vice President, Pacific Rim Objective Measurement Society The Education University of Hong Kong	SR.C.3.08 / SR.C.3.09
4:55 PM	End of Day 2	

Concurrent Sessions

Day 1 Concurrent Session I

Day 1 Concurrent Session I (11:00 am – 12:00 pm) *60 minutes*

Theme: Distinguished Student Scholarship Presentation

Venue: SR.C.3.10

Chair: Dr Mohd Zali Mohd Nor

[PROMS2025-IN011] Identifying biases occurred among Indonesian university lecturers in assessing English essays: An MFRM analysis
Author(s): Muhammad Affan Ramadhana, Bambang Sumintono & Zulfa Sakhiyya

[PROMS2025-MY001] Fair or fickle? The tug-of-war between objectivity and teacher-student bias in speaking assessments
Author(s): Muhamad Firdaus Mohd Noh, Mohd Effendi @ Ewan Mohd Matore, Nur Ainil Sulaiman

Theme: Measurement Theory & Practice

Venue: SR.C.3.15

Chair: Dr Chia-Ling Hsu

[PROMS2025-TW005] Recovering person, item and dispersion parameters in the extended continuous Rating scale model
Author(s): Yeh-Tai Chou, Yao-Ting Sung, Pin-Hsun Song

[PROMS2025-AU002] Optimizing test length in assessments through adaptive simulation
Author(s): Xiaoxun Sun

[PROMS2025-HK004] Variable-length multistage adaptive testing design
Author(s): Chia-Ling Hsu

Theme: Instrument Development & Validation

Venue: SR.C.3.14

Chair: Dr Jonathan Barcelo

[PROMS2025-SG004] Developing and validating a generative AI literacy scale in postgraduates' academic writing
Author(s): Yu Liu, Shaoyan Zou

[PROMS2025-PH003] Development and Rasch analysis of the critical thinking test in chemistry
Author(s): Jonathan Barcelo, Mark Alben Ponciano

[PROMS2025-PH002] Development and validation of the Visual Representations Test using Rasch measurement framework
Author(s): Jonathan Barcelo, Precious Lady Gine Araneta

Concurrent Sessions

Day 1 Concurrent Session II

Day 1 Concurrent Session II (1:00 pm – 2:15 pm) *75 minutes*

Theme: AI & Large Language Models

Venue: SR.C.3.10

Chair: Jade Tan

[PROMS2025-CN001] Application of LLM to optimize Q-matrix construction for cognitive diagnostic assessment in L2 reading
Author(s): Wenbo Du, Jiayi Shen, Xiaomei Ma

[PROMS2025-CN004] Exploring the moments of insight in human-AI co-creative process
Author(s): Sujie Yang, Manli Zhang, Jue Wang

[PROMS2025-SG006] Evaluating performance of large language models on university-level mathematics and psychology multiple-choice questions for adaptive learning system
Author(s): Hariz Zhen Wei Liew, Che Yee Lye

[PROMS2025-SG002] Integrating generative artificial intelligence in higher education: A pedagogical and assessment framework review
Author(s): Jade Tan, Che Yee Lye

Theme: Measurement Theory & Practice

Venue: SR.C.3.15

Chair: Dr Mohd Zali Mohd Nor

[PROMS2025-MY006] Score linking and validation in educational tests: A Rasch model study
Author(s): Zouh Fong Chieng

[PROMS2025-AU001] Equivalence of two methods for equating scores between two tests using the Rasch measurement model
Author(s): Dragana Surla, David Andrich

[PROMS2025-SG008] Identifying time-varying measurement model parameters in intensive longitudinal data using cross-classified factor model
Author(s): Ringo Moon-Ho Ho, Jie Xin Lim

[PROMS2025-SG001] Game leveling using the Rasch model
Author(s): Tzemin Chung, Mohd Zali Mohd Nor, Richard Yan, Peing Ling Loo

Theme: Teachers & Schools

Venue: SR.C.3.14

Chair: Dr Bambang Sumintono

[PROMS2025-IN010] Teachers' perception about nature of science: A Rasch model measurement analysis
Author(s): Kartimi, Siti Nadya Zynuddin, Bambang Sumintono

[PROMS2025-SG007] Pursuing a cultural understanding of distributed leadership practices among middle leaders in Singapore schools
Author(s): Simon Lim, Jonathan Goh

[PROMS2025-US001] A mixed Rasch modelling approach to investigating teacher resilience in Malaysia
Author(s): Zhi Jie Lee, Sharifah Hanizah Syed Jaafar, Esther Tan, Mei Ai Foo

[PROMS2025-US002] School bullying victimisation in Malaysia: A mixed Rasch model approach for school counselling
Author(s): Zhi Jie Lee, Mei Ai Foo, Esther Tan

Concurrent Sessions

Day 2 Concurrent Session I

Day 2 Concurrent Session I (10:15 am – 11:30 am) *75 minutes*

Theme: Instrument Development & Validation

Venue: SR.C.3.10

Chair: Assoc Prof Lyndon Lim

[PROMS2025-US003] Development of an eleven-item scale for measuring food insecurity
Author(s): Jing Li, George Engelhard Jr.

[PROMS2025-PH004] Application of Rasch analysis in the evaluation of biochemistry examination for health science students
Author(s): Jonathan Barcelo, Lloyd Allen Lorente

[PROMS2025-KR001] Development and validation of a Perceived Practical Teaching Competence Scale (PTCS) for middle school students in science classes using the partial credit model and the confirmatory factor analysis
Author(s): Sun-geun Baek, Woori Song, Yunah Kang, Byunghoon Jeon, Seojin Kim

[PROMS2025-MY008] The validation of integrating Artificial Intelligence construct for the multimodal learning framework development: A Rasch model measurement analysis
Author(s): Nurin Erdiani Mhd Fadzil, Harwati Hashim

Theme: Measurement Theory & Practice

Venue: SR.C.3.15

Chair: Prof Zi Yan

[PROMS2025-HK001] Assessing differential rater functioning with the many-facet latent space Rasch model
Author(s): Kuan-Yu Jin

[PROMS2025-HK002] Unpacking student performance in visual arts: A many-facet Rasch analysis
Author(s): Joseph Chow, Kuan-Yu Jin

[PROMS2025-TR001] Guessing as ability rather than item characteristic: A new framework for mixture Item Response Theory
Author(s): Metin Bulus

[PROMS2025-TR002] A practical guide to sample size calculations in psychometric research
Author(s): Metin Bulus

Theme: Item Bias & Test Fairness

Venue: SR.C.3.14

Chair: Prof Jue Wang

[PROMS2025-TR003] Differential item functioning analysis in PISA 2022 using Rasch trees: Finland, Turkey, and Singapore
Author(s): Enes Yavuz

[PROMS2025-TW001] Developing a revised DIF-free-then-DIF strategy to simultaneously assess uniform and nonuniform DIF
Author(s): Wei-Chia Su, Po-Hsien Hu, Ching-Lin Shih

[PROMS2025-IN003] Investigating potential differential item functioning on the Hating Adolescence Test (HAT) using Rasch model
Author(s): Nila Zaimatus Septiana, Intan Nuyulis Naeni Puspitasari, Ummiy Fauziyah Laili, Choirul Annisa, Agus Miftakus Surur, Rizqona Maharani, Suharni, Choiru Ummatin

[PROMS2025-KU001] Evaluating the fairness of a high-stakes college entrance exam in Kuwait: A Rasch model application
Author(s): Fajer Shamsaldeen, Jue Wang, Soyeon Ahn

Concurrent Sessions

Day 2 Concurrent Session II

Day 2 Concurrent Session II (11:35 am – 12:35 pm) *60 minutes*

Theme: Health & Well-being

Venue: SR.C.3.10

Chair: Prof Quan Zhang

[PROMS2025-IN001] The Indonesian Food Neophobia Scale revisited: A two-year Rasch-based study and differences in family eating habits
Author(s): Itsar Bolo Rangka

[PROMS2025-MY002] Validation of a burnout assessment tool for healthcare workers: A psychometric approach using Rasch modelling and exploratory factor analysis
Author(s): Suriya Kumareswaran Vallasamy, Rosnah Ismail

[PROMS2025-IN013] Development of a web-based ESQ assessment tool using Rasch model analysis for holistic psychological well-being
Author(s): Basma Tania, Kukuh Setyo Pambudi, Jati Fatmawiyati, Iffat Maimunah, Wildana Wargadinata, Tutut Chusniyah, Mochammad Said, Muhammad Izzudin Haq, Syabiilah Azzahroh Widyatmoko Putri, Habil Abyad

Theme: Inclusivity & Learning Needs

Venue: SR.C.3.15

Chair: Sharyfah Fitriya

[PROMS2025-HK003] Bridging theories and practice of inclusive assessment: A systematic review of frameworks and measures
Author(s): Jiaying Chen, Jiahe Gu, Zi Yan

[PROMS2025-CN005] Exploring students' sense of belonging in STEM colleges: A many-facet Rasch model approach
Author(s): Yingying Zhang, Yang Yang, Manli Zhang, Jue Wang

[PROMS2025-SG003] Towards a framework of multilevel analysis of student- and teacher-level factors influencing dyslexic students' reading performances
Author(s): Sharyfah Fitriya, Che Yee Lye

Theme: Continuous Education & Workplace Competency

Venue: SR.C.3.14

Chair: Dr Mei Teng Ling

[PROMS2025-MY003] Automated PPKI application system via Google forms and WhatsApp: Development and evaluation using Rasch measurement model
Author(s): Mei Teng Ling, Nur Hanini Anne Abdullah, Felicia Suling Emang

[PROMS2025-SG005] Standards-aligned authentic assessment in pharmacy technician education
Author(s): Yin Ni Annie Ng, Cheng Keat Tan

[PROMS2025-MY005] Expert validation of the C-A-RE module for sandwich generation workers using many-facet Rasch analysis
Author(s): Rozita Jayus, Aqeel Khan, Adibah Abdul Latiff, Mastura Mahfar, Nornazira Binti Suhairom, Siti Aminah

Concurrent Sessions

Day 2 Concurrent Session III

Day 2 Concurrent Session III (1:35 pm – 2:35 pm) *60 minutes*

Theme: Math

Venue: SR.C.3.10

Chair: Jade Tan

[PROMS2025-TW003] Development of a multidimensional mathematical competence adaptive test: Item bank construction, simulation, and empirical analysis
Author(s): Yu-Chun Lien, Yao-Ting Sung, Wei-Hung Yang

[PROMS2025-CN006] Generating math word problems aligned with pupil ability and item difficulty
Author(s): Jie Wang, Xinguo Yu

[PROMS2025-IN008] Evaluating students' performance on cryptarithms: Item analysis from a pilot study
Author(s): Elizar, Anwar, Ayu Mastura

Theme: Language & Multiple Intelligence

Venue: SR.C.3.15

Chair: Dr Chia-Ling Hsu

[PROMS2025-TW002] Development and validation of a framework for assessing linguistic competencies in senior-year Chinese majors
Author(s): Suet Ching Soon, Chia-Ling Hsu

[PROMS2025-TW004] Predicting IRT-based word difficulty using deep neural networks: A semantic feature-based approach
Author(s): Wei-Hung Yang, Yao-Ting Sung, Yu-Chun Lien, Chia-Hsin Chen

[PROMS2025-CN007] Multiple intelligence assessment for rural primary students: Promoting equity through gamified-designed and non-graded inventory
Author(s): Kaixin Liang, Wen Qin, Rou Chen, Ziqi Li, Xiaomin Mai

Theme: 21st Century Competencies & Skills

Venue: SR.C.3.14

Chair: Dr Jonathan Barcelo

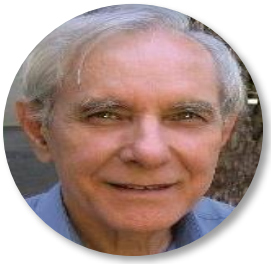
[PROMS2025-PH001] Path analysis of critical thinking in chemistry informed by the Rasch measurement framework
Author(s): Jonathan Barcelo

[PROMS2025-MY007] Psychometric validation of a 21st century skills instrument in a design thinking context among final year polytechnic students using the Rasch measurement model
Author(s): Aede Hatib Musta'amal, Nor Aisyah Che Derasid, Mohd Safarin Nordin, Nornazira Suhairom, Rozita Jayus

[PROMS2025-IN005] The expert judgement validation of Student Growth Mindset Scale (SGMS) using Many Facet-Rasch Measurement (MFRM)
Author(s): Ma'rifatin Indah Kholili, Nandang Rusmana, Ahman, Nandang Budiman, Rahmi Ramadhani

Pre-conference Workshops

21 July 2025, Monday, 9:30 AM – 12:30 PM



Emeritus Professor David Andrich
The University of Western Australia

Workshop Title: Adapting the Rasch Meta-metre of Growth for Variables in the Social Sciences

Workshop Synopsis

Georg Rasch is well-known for applying his principle of invariant comparisons to provide interval level measurements. He also studied physiological growth applying the same principle. His approach identifies a transformation of time, called the *meta-metre,* within which every individual's rate of growth is linear and therefore also invariant. Adapting Rasch's approach to growth for social science variables requires interval level measurements. For studies of growth, the instruments of measurement generally require a linked design of increasing item difficulty that ensures items are successively aligned to each individual's stage of growth, perhaps across 10 years. The workshop has two parts: first, it shows the invariant properties of items necessary for linked designs and an efficient approach to diagnosing any lack of invariance; and second, it introduces an adaptation of Rasch's approach to estimating a meta-metre of growth for social science variables with illustrations from simulated and real data.

Speaker Profile

David Andrich is Emeritus Professor of Education, The University of Western Australia. His interests are in educational, psychological and social measurement in general, where he is best known for his work on Rasch measurement theory, including its applications through software development. In 1990, he was elected *Fellow of the Academy of Social Sciences of Australia* for his contributions to measurement in the social sciences. He has published in Educational, Psychological, Sociological, Statistical and more recently, Physics measurement journals. In addition to many articles, he is the author of *Rasch Models for Measurement* (Sage, 1998) and coauthor of the *A Course in Rasch Measurement Theory: Measuring in the Educational, Social and Health Sciences* (Springer, 2019). His more recent work has been on the study of growth in attainment tests reflected in the publication Andrich, D., Marais, I. and Sappl, S. (2023) *Rasch Meta-Metres of Growth for Some Intelligence and Attainment Tests*. Springer, Singapore.

Pre-conference Workshops

21 July 2025, Monday, 9:30 AM – 12:30 PM

Dr Rassoul Sadeghi
Australian Curriculum, Assessment and Reporting Authority

Workshop Title: Item Banking and Adaptive Testing

Workshop Synopsis

This workshop is tailored for all participants eager to delve into advanced concepts in educational measurement. It offers a comprehensive introduction to item banking, emphasising the design and management of test items enriched with detailed metadata. Participants will explore how adaptive testing leverages item banks to tailor assessments dynamically, aligning item difficulty with test-takers' abilities to enhance accuracy and efficiency. The workshop will underscore the pivotal role of item banking as the foundation for effective adaptive testing, ensuring both precision and fairness. Furthermore, it will cover test equating using the Rasch measurement model, illustrated through examples from prominent large-scale assessments like NAPLAN. Ideal for participants aiming to advance their expertise, this session bridges theory and practice in modern assessment design.

Speaker Profile

Dr Rassoul Sadeghi is a Lead Psychometrician at Australian Curriculum, Assessment and Reporting Authority (ACARA). He has been with ACARA since 2015. Before joining ACARA, he worked as a Senior Psychometrician at Educational Assessment Australia (EAA). In his current position, he is responsible for psychometric aspects of the National Assessment Program in Numeracy and Literacy (NAPLAN online). He has more than 25 years of expertise as a Psychometrician, specialising in the following areas:

- Management and analysis of large and complex data sets
- Test equating and scaling using the Rasch measurement model
- Development and maintenance of item bank
- Designing both ‘low stake’ and ‘high stake’ online assessment programs
- Designing and implementation of adaptive testing (CAT and MST)
- Resolving measurement issues emerging from live testing situations

Pre-conference Workshops

21 July 2025, Monday, 1:30 PM – 4:30 PM

Professor Jue Wang
University of Science and Technology of China

Workshop Title: Rasch Measurement for Rater-mediated Assessment

Workshop Synopsis

Rating scales are widely used for human judgments across the social, behavioral, and health sciences, from high-stakes performance assessments in education and personnel evaluations to functional assessments in medical research. This workshop applies principles of invariant measurement and lens models from cognitive psychology to examine judgment processes in rater-mediated assessments, focusing on creating, evaluating, and maintaining invariant systems (Engelhard & Wang, 2024). We introduce rater-mediated assessments such as performance assessments, demonstrating how Rasch models can provide item-invariant person measurement and person-invariant item calibration. Building on these foundations, the workshop explores the Many-Facet Rasch Model for developing robust performance assessments, illustrated by large-scale writing examples. Participants are encouraged to bring their own data for hands-on analysis and discussion, and will gain practical strategies for practices on using rating scales. Throughout the workshop, the Facets software (Linacre, 2024) will exemplify these principles, supporting the implementation of rating scales that yield reliable and meaningful human judgments.

Speaker Profile

Jue Wang, PhD, is currently a professor in Department of Psychology at The University of Science and Technology of China. Dr Wang received her Ph.D. in Quantitative Methodology Program under Educational Psychology at The University of Georgia, and previously worked in Research, Measurement & Evaluation Program at The University of Miami. Her research focuses on examining rater effects in rater-mediated assessments, such as writing assessments and creativity assessments, using Rasch measurement models and unfolding models. She has published in peer-reviewed journals including *Educational Psychology Review*, *Psychology of Aesthetics, Creativity, and the Arts*, *Educational and Psychological Measurement*, *Journal of Educational Measurement*, and *Assessing Writing*. Dr Wang has co-authored two books (with Professor George Engelhard): *Rasch models for solving measurement problems: Invariant measurement in the social sciences* published by Sage as part of Quantitative Applications in the Social Sciences (QASS) series, and *Invariant Measurement: Using Rasch Models in the Social, Behavioral, and Health Sciences (2nd edition)* by Routledge.

Pre-conference Workshops

21 July 2025, Monday, 1:30 PM – 4:30 PM

Professor Quan Zhang
Jiaxing University, Zhejiang, China
The World Sports University, Macau SAR, China

Workshop Title: Rasch-GZ: Item Analysis and Test Equating in the Age of AI

Workshop Synopsis

The purpose of this workshop is to introduce important applications of Rasch model to language testing: Item Analysis and Test Equating via Rasch-GZ. The workshop falls into three parts.

(1) Introduction to Rasch-GZ;
(2) Demonstration of Rasch-GZ and
(3) Q & A.

Participants needn't have particular psychometric competence. Just bring their own laptop computers to download Rasch-GZ, free of charge, to install in their computers for future use. No other pre-requisites or requirements.

For more detailed features of Rasch-GZ, please click the following link: https://doi.org/10.2991/978-94-6463-494-5_25

For more info about PROMS conference, you are welcome to visit https://atlantis-press.com/proceedings/proms-23.

Speaker Profile

Professor Zhang Quan, PhD, a PROMS board member, a Professor of Jiaxing University/the World Sports University, Macau, China. Since 1989, he has been involved in test equating for large-scale, high-stakes language assessments. He currently serves as China representative of PROMS, editor of PROMS conference Proceedings, and reviewer of several prestigious academic journals. Ever since 2012, he has been organizing or helping organize PROMS conferences. During the global fighting against COVID-19 pandemic period, he organised a small team of qualified computer engineers and testing professionals to have developed Rasch-GZ, the Rasch-based software ad hoc for item analysis and test equating.

PROMS2025 Highlights

22 July 2025, Tuesday, 10:00 AM – 10:50 AM

Professor George Engelhard Jr.
President, Pacific Rim Objective Measurement Society
University of Georgia

Presidential Address: Measurement, Explanation, and Invariance: Next Generation Invariant Measurement

Presentation Abstract

Science is built upon relationships between measurement, explanation, and invariance. My address emphasizes that the principles of invariance are essential to measurement and scientific objectivity. Rasch measurement theory highlights the role of objectivity in creating and using invariant scales. In addition to this connection between measurement and invariance, science relies on discovering explanations that reflect stable relationships between variables. We seek invariant relationships across various conditions. Explanatory Item Response Models (EIRMs) can provide an approach for linking measurement and explanations based on principles of invariance. The essential principles for the next generation of invariant measurement include the following:

- Constructs should be unidimensional and defined by latent variables,
- the measurement of persons should be invariant across items,
- item calibrations should be independent of specific persons, and finally,
- structural analyses should seek invariance in relationships among variables.

Next generation invariant measurement should provide the integration of measurement theory with explanatory models to deepen our understanding, and to discover invariant relationships in the human sciences.

Speaker Profile

Professor George Engelhard, Jr., PhD, is at The University of Georgia. Professor Engelhard received his PhD in 1985 from The University of Chicago in the MESA Program (measurement, evaluation, and statistical analysis). While he was at The University of Chicago, he worked closely with Professors Ben Bloom and Ben Wright. Professor Engelhard is the author of several books including his latest with Dr Jue Wang: *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences (2nd Edition)*. He is currently president of the Pacific Rim Objective Measurement Society. He is a fellow of the American Educational Research Association.

PROMS2025 Highlights

22 July 2025, Tuesday, 2:25 PM – 3:15 PM

Professor Jue Wang
University of Science and Technology of China

Keynote Presentation: Addressing Human Scoring in Subjective Creativity Assessments

Presentation Abstract

This talk provides an in-depth examination of human-based scoring methods for subjective creativity assessments (SCA). It begins with an overview of the PISA 2022 framework on creative thinking, followed by two empirical studies exploring various scoring techniques. The first study illustrates the use of a partial credit model to identify rater effects and rating scale malfunctioning in expert scoring methods, while also examining how rater judgments are influenced by features of creative responses. The second study explores peer scoring method as an alternative approach for classroom creativity assessments. This talk primarily emphasizes the complexities and challenges of human judgment in evaluating creativity, offering insights into improving the reliability and validity of creativity assessments, as well as the development of automated scoring methods.

Speaker Profile

Jue Wang, PhD, is currently a professor in Department of Psychology at The University of Science and Technology of China. Dr Wang received her Ph.D. in Quantitative Methodology Program under Educational Psychology at The University of Georgia, and previously worked in Research, Measurement & Evaluation Program at The University of Miami. Her research focuses on examining rater effects in rater-mediated assessments, such as writing assessments and creativity assessments, using Rasch measurement models and unfolding models. She has published in peer-reviewed journals including *Educational Psychology Review*, *Psychology of Aesthetics, Creativity, and the Arts*, *Educational and Psychological Measurement*, *Journal of Educational Measurement*, and *Assessing Writing*. Dr Wang has co-authored two books (with Professor George Engelhard): *Rasch models for solving measurement problems: Invariant measurement in the social sciences* published by Sage as part of Quantitative Applications in the Social Sciences (QASS) series, and *Invariant Measurement: Using Rasch Models in the Social, Behavioral, and Health Sciences (2nd edition)* by Routledge.

PROMS2025 Highlights

22 July 2025, Tuesday, 3:45 PM – 4:55 PM

Plenary Roundtable Discussion

Title: The Road Ahead: Future Challenges and Opportunities in Objective Measurement

Presentation Abstract

This plenary session examines how Artificial Intelligence (AI) can enhance Rasch-based assessment and measurement. We will explore AI's potential to improve large-scale assessments through automated item generation and adaptive testing, leading to better scalability, efficiency and precision. The discussion will address ethical considerations, particularly algorithmic bias, and the importance of maintaining fair practices. We emphasise the need for collaboration between psychometricians, researchers and AI specialists in this interdisciplinary field. Key topics include the evolving skillsets needed for AI integration and the challenges of automating Rasch analyses without compromising measurement integrity. The session aims to initiate dialogue about the future of AI in Rasch measurement, working towards responsible innovation that enhances our understanding of human development and performance while effectively managing the complexities of AI implementation.

Speakers' Profiles

Professor George Engelhard Jr.
University of Georgia

Professor George Engelhard, Jr., PhD, is at The University of Georgia. Professor Engelhard received his PhD in 1985 from The University of Chicago in the MESA Program (measurement, evaluation, and statistical analysis). While he was at The University of Chicago, he worked closely with Professors Ben Bloom and Ben Wright. Professor Engelhard is the author of several books including his latest with Dr Jue Wang: *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences (2nd Edition)*. He is currently president of the Pacific Rim Objective Measurement Society. He is a fellow of the American Educational Research Association.

PROMS2025 Highlights

Professor Zi Yan
The Education University of Hong Kong

Professor Zi Yan, PhD, is a Hong Kong RGC Senior Research Fellow and serves as the Head of the Department of Curriculum and Instruction at The Education University of Hong Kong. Additionally, he holds the title of Honorary Professor at the Centre for Research in Assessment and Digital Learning at Deakin University. His research and publications primarily focus on two areas: educational assessment izzn both school and higher education contexts, with a particular emphasis on student self-assessment, and Rasch measurement, specifically its application in educational and psychological research. He is also the co-author of the book 'Applying the Rasch model: Fundamental measurement in the human sciences (4th ed.)'.

Professor Quan Zhang
Jiaxing University, Zhejiang, China
The World Sports University, Macau SAR, China

Professor Zhang Quan, PhD, a PROMS board member, a Professor of Jiaxing University/the World Sports University, Macau, China. Since 1989, he has been involved in test equating for large-scale, high-stakes language assessments. He currently serves as China representative of PROMS, editor of PROMS conference Proceedings, and reviewer of several prestigious academic journals. Ever since 2012, he has been organizing or helping organize PROMS conferences. During the global fighting against COVID-19 pandemic period, he organised a small team of qualified computer engineers and testing professionals to have developed Rasch-GZ, the Rasch-based software ad hoc for item analysis and test equating.

Professor Guanzhong Luo
South China Normal University
Jiangxi Normal University of China

Professor Guanzhong Luo holds a doctoral degree in Mathematical Psychology and has served as the Director of Assessment Technology and Research at the HKEAA for nearly 15 years. He now holds a professorship at South China Normal University and Jiangxi Normal University of China. Professor Luo's research focuses on psychometric models and parameter estimation algorithms for achievement and attitude measurements, and his publications and computer programs for test data analysis are widely used globally.

PROMS2025 Highlights

Associate Professor
Jeffrey Durand
Toyo Gakuen University

Jeffrey Durand is an associate professor in the Faculty of Global Communication at Toyo Gakuen University in Tokyo, Japan. His background is in language education and language testing, especially rater-mediated assessment using many-facet Rasch measurement. He also has research interests in motivation, student study abroad, global mindset, and other intercultural issues. Finally, he teaches a number of courses related to global issues and globalisation.

Dr Mohd Zali Mohd Nor
Newstar Agencies Sdn. Bhd.

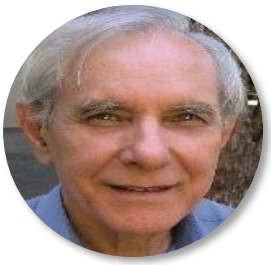
Dr Mohd Zali Mohd Nor is an I.T. Manager in a Malaysian shipping agency, managing a team of analysts and developers to develop and maintain world-wide enterprise shipping solutions. He received his B.Sc. in Mathematics and Computing from The University of Michigan, Ann Arbor, MI, USA, in 1988, Master of Management in I.T. from Universiti Putra Malaysia in 2005, and PhD in Management Information System from Universiti Putra Malaysia in 2012. His involvement in Psychometric and Rasch Measurement Models started in 2009, specialising in Rating Scale, Partial Credit and Multi-Facet Rasch models (MFRM). He is currently active in providing trainings and academic consultations on Rasch measurement and has assisted many postgraduate students from various local and international universities on research methodology and Rasch analyses. He has also served as psychometrician in several assessment projects with the Department of Education Malaysia, Fire and Rescue Department and National Child Development Research Center (NCDRC). He is currently the Vice President of the Pacific Rim Objective Measurement Symposium (PROMS), a committee member of Malaysian Psychometric Association (MPA) and Vice President of myRasch.

Moderator
Dr Iris Lee
Singapore Ministry of Education

Dr Iris Lee is an education professional with a PhD completed in 2007. Her career spans teaching in Singapore and Hong Kong, followed by roles at Singapore's Ministry of Education (MOE) and a secondment to NIE/NTU. A longstanding member of the Pacific Rim Objective Measurement Society (PROMS), she first attended their conference in 2007, engaging with experts like Prof Mike Linacre and the late Prof Wang Wenchong. In her current role at MOE, Dr Iris Lee specialises in survey development, data analysis, and research-related tasks, applying her expertise to shape educational policies and practices in Singapore's education system.

PROMS2025 Highlights

23 July 2025, Wednesday, 9:00 AM – 9:50 AM



Emeritus Professor David Andrich
The University of Western Australia

Distinguished Scholar Presentation: Rasch Meta-metres of Growth in Reading and Mathematics Attainment at Both Population and Individual Levels

Presentation Abstract

Although hardly known, Georg Rasch had an approach to studying growth based on the principle of *invariant comparisons*, a principle for which he is well known with his models for measurement. The approach identifies a non-linear function of time, called a *meta-metre*, which governs the growth of all individuals of a population. Then within the meta-metre, each individual's rate of growth is linear and *invariant*, thus permitting comparisons among individuals using standard statistical procedures. This address illustrates the approach with the educationally important variables of reading and mathematics attainment tests from two longitudinal studies. Each of the meta-metres show early rapid, decelerating growth, with noticeably different rates of growth among sub-populations. Decelerating growth is also related to the common grade scale, showing that any grade difference between groups in the early years invariably increases in later years. This increase has implications for interventions for groups at risk in their attainments.

Speaker Profile

David Andrich is Emeritus Professor of Education, The University of Western Australia. His interests are in educational, psychological and social measurement in general, where he is best known for his work on Rasch measurement theory, including its applications through software development. In 1990, he was elected *Fellow of the Academy of Social Sciences of Australia* for his contributions to measurement in the social sciences. He has published in Educational, Psychological, Sociological, Statistical and more recently, Physics measurement journals. In addition to many articles, he is the author of *Rasch Models for Measurement* (Sage, 1998) and coauthor of the *A Course in Rasch Measurement Theory: Measuring in the Educational, Social and Health Sciences* (Springer, 2019). His more recent work has been on the study of growth in attainment tests reflected in the publication Andrich, D., Marais, I. and Sappl, S. (2023) *Rasch Meta-Metres of Growth for Some Intelligence and Attainment Tests*. Springer, Singapore.

PROMS2025 Highlights

23 July 2025, Wednesday, 2:40 PM – 3:30 PM

Associate Professor Vahid Aryadoust
National Institute of Education, Nanyang Technological University, Singapore

Keynote Presentation: What Can Generative AI Do for Assessing Listening and Speaking Skills

Presentation Abstract

In this talk, I draw on the forthcoming book *Assessing Listening in the Age of Generative Artificial Intelligence* (Aryadoust, 2025) to discuss how AI technologies can be used for listening and speaking assessment. These technologies include foundation models, text-to-speech, and automated speech recognition. Foundation models, such as large language models, are trained on extensive datasets, which enables them to address a wide range of tasks. Text-to-speech and automated speech recognition systems generally function as more specialized components for developing AI-based assessments. When these AI-driven technologies are combined, they can create robust and scalable frameworks for designing language assessment tools and evaluating oral interaction competence. These innovations have great potential for educators, researchers, and practitioners, as they offer a set of integrated tools to address the needs of listening and speaking assessment in a world increasingly shaped by generative AI.

Speaker Profile

Dr Vahid Aryadoust is an Associate Professor of Language Assessment at the National Institute of Education, Nanyang Technological University (NTU), Singapore. He teaches graduate and doctoral courses on generative artificial intelligence in language assessment and research methodology, while serving as the Research Program Leader in his department and supervising Master's and PhD students. His research spans sensor technologies such as eye tracking and neuroimaging in language assessment, generative AI applications, meta-analysis, and scientometrics, with extensive publications in these areas. A multi-award-winning scholar, Dr Aryadoust and his team received the International Language Testing Association's (ILTA) Best Article Award in 2024 for their groundbreaking paper on the application of sensor technologies in listening assessment. He also runs a YouTube channel, *Statistics and Theory*, which promotes open access to knowledge and science (https://m.youtube.com/@VahidAryadoust).

PROMS2025 Highlights

23 July 2025, Wednesday, 3:55 PM – 4:45 PM

Dr Che Yee Lye
Chair, Pacific Rim Objective Measurement Symposium 2025
Singapore University of Social Sciences

Keynote Presentation: Refocusing Educational Measurement: Understanding Student Learning through Adaptive Learning

Presentation Abstract

Adaptive learning is a promising technology that is transforming higher education by personalising instruction to meet the diverse needs of adult learners. This technology leverages data-driven algorithms and continuous assessment to tailor content and pace to individual needs, enhancing learning outcomes. Traditional methods of assessment and measurement tend to focus on the starting and ending performance points, neglecting progression of a learner's development. Furthermore, a significant tension exists between using assessment data for institutional accountability and improvement in learning and teaching. I will argue that our adaptive learning pedagogy provides an opportunity to adopt a broader approach to measuring student progress by prioritising learning processes and growth, and integrating human-machine collaboration in both the development of adaptive learning content and the nuanced interpretation of learner data.

Speaker Profile

Che Yee Lye, PhD, is currently Senior Lecturer with Singapore University of Social Sciences (SUSS) where she teaches assessment-related courses and conducts professional development training on AI, measurement and assessment for SUSS faculty and associates. Dr Lye received her PhD (Education) from The University of Adelaide. Her current research focuses on adaptive testing and learning using Rasch measurement models, AI and assessment, as well as language curriculum and assessment. Before joining SUSS, she served as a Senior Curriculum Developer at the United Chinese School Committees' Association of Malaysia, where she researched English language curriculum and evaluation, published educational materials, and conducted teacher training. She was also a Research Specialist at the Singapore Examinations and Association Board, focussing on assessment for learning as well as computerised and multistage adaptive testing for primary school Mathematics and English. In her current role with SUSS, she leads the AdLeS Research Group (ARG) in developing the Adaptive Learning System (AdLeS). Since 2021, AdLeS has served over 2,700 students, enabling instructors to monitor progress and identify students needing support, and providing students with meaningful feedback on learning. She is also Chair of Special Interest Group on Generative AI & Learning, Teaching and Assessment (SIG-AILTA) and co-manages https://sigailta.com, a blog dedicated to sharing ideas and resources on pedagogical and assessment practices in the era of generative AI.

Identifying biases occurred among Indonesian university lecturers in assessing English essays: An MFRM analysis

by Muhammad Affan Ramadhana | Universitas Islam Internasional Indonesia

Bambang Sumintono | Universitas Islam Internasional Indonesia

Zulfa Sakhiyya | Universitas Negeri Semarang

Abstract ID: PROMS2025-IN011

Presenter Name(s): Muhammad Affan Ramadhana

Abstract:

Assessment of writing in English as a Foreign Language (EFL) context often utilize standard analytical rubric. However, writing assessment remains a human judgment process susceptible to inconsistency and bias despite using standardized rubrics. To further explore to that issue, the present study aims to utilize Many-Facet Rasch Measurement (MFRM) analysis to identify the bias/interaction from raters' background towards rubric criteria in English essay assessment context. This study includes 36 Indonesian university lecturers with postgraduate degrees in English language education, linguistics, or literature studies. They assessed English essays by utilizing the ESL Composition Profile analytical rubric. Prior to the assessment, they were invited to complete rater training to familiarize themselves with the rubric. The rating data was analyzed using the Many-Facet Rasch Measurement model, with bias/interaction model between rater facet and criteria facet. Additionally, several dummy facets are also created to examine the biases/interactions between rater backgrounds and assessment criteria. The analysis shows that raters with PhDs are notably more severe in scoring Content and Language Use, but more lenient in Organization, Vocabulary, and Mechanics. Raters from Linguistics or Literature study background are the most lenient on Content but become most severe on Language Use. Moreover, raters who are 'Associate Professors' show high severity on Vocabulary and Organization, but exceptionally lenient on Content and Mechanics. In terms of gender, male raters show more severity on Content and Organization but become extremely lenient on Language Use. In contrast, female raters show slight leniency on Content and Organization, but gradually more severe on Vocabulary, Language Use, and Mechanics. The bias/interaction analysis suggests gender-based differences in rubric interpretation. From the initial findings in this study, it can be concluded that there is noticeable scoring differences among raters based on several factors such as academic qualifications, field of study, professional rank, and gender. However, to understand more about this pattern we still need further studies to look deeper at the significance and size of these biases.

Keywords: Writing assessment, rating bias, rating behavior, rater background, distal facets

Fair or fickle? The tug-of-war between objectivity and teacher-student bias in speaking assessments

by Muhamad Firdaus Mohd Noh | Universiti Kebangsaan Malaysia

Mohd Effendi @ Ewan Mohd Matore | Universiti Kebangsaan Malaysia

Nur Ainil Sulaiman | Universiti Kebangsaan Malaysia

Abstract ID: PROMS2025-MY001

Presenter Name(s): Muhamad Firdaus Mohd Noh

Abstract:

Despite existing research on biases in teacher rating, there remains a critical gap in understanding how student factors influence rating consistency. The study aims to determine bias interaction between teachers and students during the marking of a speaking test. The conceptual framework is informed by Lens Model (Brunswik, 1952) that conceptualizes the assessment process through three key components: the construct (speaking proficiency), the cues (teachers, students, items, and domains), and the evaluation (scores). The research question is to what extent do teacher-student interactions influence bias in the marking of English language speaking proficiency? The null hypothesis states that there is no statistically significant bias interaction between teachers and students. The study recruited a sample of 164 English teachers to mark the speaking test responses of 30 students across three task types: background interview, storytelling, and discussion. Teachers underwent standardized training to enhance rating consistency. A linked rating design was employed, ensuring systematic connections among key facets, thereby enabling a robust analysis through the Many-Facet Rasch Measurement (MFRM) model. The analysis of teacher-student interactions identified 982 instances, with 253 (25.76%) exceeding the t-value threshold of ±2.00, indicating statistically significant bias. Severity biases were slightly more frequent (53.17%) than leniency biases (46.83%). Analyzed based on students' ability levels, mid-range ability students exhibited the highest proportion of bias (45.63%), followed by high-ability students (29.76%) and low-ability students (24.6%). This pattern implies that teachers may find it more challenging to assess students whose proficiency is not clearly distinguishable, leading to greater variability in their scoring. Conversely, low- and high-ability students may exhibit more distinct language performance. The uneven distribution of bias across ability levels raises concerns about the fairness of speaking assessments, as mid-range ability students who represent a significant portion of the student population are more prone to inconsistent ratings. This study highlights the need for continuous professional development and structured rating rubrics to minimize bias to ensure equitable evaluations. While the study is limited to secondary school English teachers in Malaysia, future research should explore rating bias across different subjects and educational levels to enhance the generalizability of findings.

Keywords: Rating bias, teacher rating, multi-facet Rasch measurement, speaking test, language testing

Equivalence of two methods for equating scores between two tests using the Rasch measurement model

by Dragana Surla | School Curriculum and Standards Authority, Western Australia

David Andrich | The University of Western Australia

Abstract ID: PROMS2025-AU001

Presenter Name(s): Dragana Surla

Abstract:

Equating test scores has a major role in educational measurement. The Rasch measurement model provides two methods for equating tests using only test scores available from common persons. The contrast between the two methods is that the first uses only the total score on two tests, while the second method involves proficiency values on a common scale. Both methods involve first estimating the test parameters from the total sample of persons. From method 1, for every total score on the two tests, the two equated scores on the two tests are the respective expected values given the total score. From method 2, for any real proficiency value on the scale for a score on a specific test, the equated score on any other test is the expected value (theoretical mean) given the proficiency on the specific test. As shown in the example, the two methods give virtually identical equated scores. There are two disadvantages in generalising the first method to more than two tests. The first is that it involves symmetric functions, and with a large range of scores, such as 0 to 100, it is virtually impossible to implement. On the other hand, given parameter estimates of test from the second method the equated expected values are obtained readily.The second disadvantage is that the equated values are generally real values for both tests and the equivalent of one test to an integer value of another test is approximated. The second method can take advantage of the sufficiency of each test score for the proficiency estimate of that score for any chosen test. Then the equated score on any other test, of the integer score of a specific test, is simply the expected value given the proficiency estimate of that score for that chosen test. In addition to an example of equated scores of two tests by the two methods, to illustrate the advantage of the second method, an example of equated scores of six tests using the second method will be presented.

Keywords: Test equating, score equating, Rasch model equating

Optimizing test length in assessments through adaptive simulation

by Xiaoxun Sun | Australian Council for Educational Research

Abstract ID: PROMS2025-AU002

Presenter Name(s): Xiaoxun Sun

Abstract:

This study investigates the feasibility of reducing the number of items in assessments without compromising measurement precision or classification accuracy. Grounded in item response theory (IRT), the research applies adaptive testing principles to fixed-form assessments. The objective is to determine whether shorter tests can deliver psychometric results comparable to full-length versions, thereby improving efficiency and reducing the cognitive burden on examinees. Response data were used from one form each of a 60-item Numeracy and Literacy assessment. Computerized adaptive testing (CAT) simulations were conducted, tracking ability estimates and standard errors of measurement (SEMs) across item positions to identify the point of ability stabilization. The impact of item ordering on convergence and classification consistency was also examined. Results show that for both Literacy and Numeracy forms, ability estimates stabilized within the first 40 items, with SEMs remaining within acceptable thresholds. The correlations between ability estimates from the 40-item tests and the full-length tests were 0.88 (Numeracy) and 0.91 (Literacy). Pass/fail classification showed 95.1% (Numeracy) and 92.3% (Literacy) agreement. Furthermore, rearranging item order had some impact: for the Numeracy form, the number of items could be reduced further to approximately 30, consistent with expectations under CAT when items are ordered by difficulty, allowing faster convergence with less fluctuation. Our findings suggest that assessments can be shortened by approximately one-third without significant loss of accuracy or reliability. However, the study is currently limited to one form each for Numeracy and Literacy. Future work will extend the investigation to multiple forms to examine the consistency of these patterns. We also plan to explore profiling candidates and proposing tailored test models that accommodate diverse proficiency levels to further personalize assessment experiences.

Keywords: Test length reduction, adaptive testing simulation, measurement precision, classification accuracy, psychometric analysis

Application of LLM to optimize Q-matrix construction for cognitive diagnostic assessment in L2 reading

by Wenbo Du | Xi'an Jiaotong University

Jiayi Shen | Xi'an Jiaotong University

Xiaomei Ma | Xi'an Jiaotong University

Abstract ID: PROMS2025-CN001

Presenter Name(s): Wenbo Du, Jiayi Shen

Abstract:

Q-matrix, a core component under the framework of cognitive diagnostic assessment (CDA), specifies the relationships between test items and target cognitive skills. Traditional manually constructed Q-matrix is constrained by expert subjectivity, inefficiency, and limited robustness. To address this issue, this study, by leveraging the Deepseek large language model, proposes a human-AI collaborative framework to automate and refine Q-matrix construction for assessing L2 reading inferential skills. Two research questions are investigated:1) To what extent can an LLM-generated Q-matrix achieve comparable or superior diagnostic capacity to manually constructed Q-matrices? 2) Does a hybrid human-AI revised Q-matrix enhance the diagnostic capacity over purely automated or manual approaches? Following the procedure of CDA, two inputs are required, i.e., Q-matrices and test response data. To this end, five Q-matrices were constructed from different sources, including researcher (Qmat-R), experts (Qmat-E), students (Qmat-S), Deepseek-driven (Qmat-DS), and human-AI revised version (Qmat-DS-H). A sample of 1083 students' test response data of an online diagnostic reading inferential test were utilized in the CDA estimation process. G-DINA model was then applied to check the diagnostic capacity of the above mentioned Q-matrices based on two types of indices: the model-data fit statistics and classification accuracy. The CDA estimation was conducted using G-DINA package (version 2.9.4) embedded in R studio. Results showed that Deepseek-generated Q-matrix (Qmat-DS) generally demonstrated superior model-data fit and comparable classification accuracy to the three manually constructed Q-matrices (Qmat-R, Qmat-E and Qmat-S). It also showed a relatively high skill-level classification accuracy (over .8) across all eight tested reading skills. Its test-level classification accuracy, however, was slightly lower than the cut-off value .7. This deficiency was largely enhanced by human-AI revised Q-matrix (Qmat-DS-H). The model-data fit and classification accuracy of Qmat-DS-H surpassed those of Qmat-DS and manually constructed Q-matrices. To sum up, this study demonstrates the viability of LLMs in optimizing Q-matrix construction, with human-AI collaboration mitigating manual limitations. The framework enhances efficiency while maintaining interpretability, offering a paradigm for scalable cognitive diagnostic tool development. Limitations include the model's dependency on high-quality prompt engineering and its untested generalizability to other language skills.

Keywords: Q-matrix, cognitive diagnostic assessment, large language model, L2 reading, human-AI collaboration

Exploring the moments of insight in human-AI co-creative process

by Sujie Yang | University of Science and Technology of China

Manli Zhang | University of Science and Technology of China

Jue Wang | University of Science and Technology of China

Abstract ID: PROMS2025-CN004

Presenter Name(s): Sujie Yang

Abstract:

In the era of generative artificial intelligence (AI), large language models such as GPT-4, DeepSeek, and Qwen increasingly collaborate with humans in creative tasks, from idea generation to problem solving (Rafner et al., 2023). Existing studies on creative process mostly focused on how human-AI interaction can enhance AI's creative output or optimizing human-AI workflows (Hitsuwari et al., 2023; Jeon et al., 2021; Rezwana & Maher, 2023). However, a critical question remains whether creativity genuinely emerges in the process of human-AI collaboration. This study investigates the bidirectional dynamics of inspiration between humans and generative AI, with a specific focus on the unique illumination stage of creative process, where sudden insights ("Eureka!" moments) catalyze novel solutions (Weisberg, 2018; Kounios & Beeman, 2014). We recruited 50 college students to solve creativity tasks through collaboration with DeepSeek, where their eye-movements were simultaneously tracked. Two science tasks were used with instructions that required participants to be as creative as possible in generating solutions, followed by an interview to report their insight moments. We first conducted a qualitative analysis of participants' report and their dialogues with DeepSeek to identify the moments of insight, based on predefined coding criteria including sudden comprehension, realization, problem reorganization or positive burst of emotion that led to a novel solution. The qualitative analysis helped define the time windows and areas of interest for the eye-tracking analyses to count fixation duration, saccadic movements, pupil dilation, scan paths (Duchowski, 2007), reflecting the sudden changes in cognitive and affective states during insight moments. We also displayed dwell time, which is the period of gaze staying within an area of interest, as heat maps (Raschke et al., 2013), to help visualize and facilitate the interpretation of insight moments. Results will be presented at the conference. This study can shed light on how moments of insight are triggered in the human-AI co-creative process and uncover the dynamics of how humans and AI mutually inspire each other to ultimately foster collaborative creativity.

Keywords: Human-AI collaboration, creative process, insight, generative AI, science tasks

Exploring students' sense of belonging in STEM colleges: A many-facet Rasch model approach

by Yingying Zhang | University of Science and Technology of China

Yang Yang | University of Science and Technology of China

Manli Zhang | University of Science and Technology of China

Jue Wang | University of Science and Technology of China

Abstract ID: PROMS2025-CN005

Presenter Name(s): Yingying Zhang, Yang Yang

Abstract:

Sense of belonging (SOB), defined as one's feeling of being personally accepted, respected, included, and supported within an environment (Goodenow & Grady, 1993), profoundly affects students' wellbeing and achievement in higher education, with both immediate and long-term effects. Existing measures of SOB often rely on self-reported surveys with ordinal Likert-scale ratings, validated through classical test theory-based approaches such as factor analysis. While these approaches often assume multidimensionality, most SOB measures to date are actually unidimensional (Dias-Broens et al., 2024). Moreover, the Likert ratings should be treated as ordinal instead of continuous as having equal intervals. This study thus addresses these gaps by developing reliable, valid, and fair measures of SOB based on Rasch measurement theory that provides sample-free calibration, measurement invariance, and item-level diagnostic precision, supporting comparable measurement across diverse higher educational settings. We created the items for measuring SOB based on the PSSM (Goodenow, 1993) and NSSE (National Survey of Student Engagement, 2020). Data responses of 1,137 STEM-major undergraduate students in a top science and technology university in China were analyzed using a many-facet Rasch model. The psychometric quality of SOB measures was evaluated through the reliability, validity and fairness arguments. We also examine the group differences on SOB measures between subpopulation groups defined by gender, grades, economic zones of residency and expected highest degrees. Results indicated a perfect reliability of separation (>0.99) for items with acceptable outfit and infit mean squares (MnSq; between 0.5 and 1.5 except for Item 6 with an outfit MnSq of 1.6). The reliability of separation for persons is generally high (0.78). Significant group differences in SOB measures were found between different groups of gender, grades, economic zones of their residency and their expected highest degrees. Certain items were found to show differential item functioning across subgroups. Detailed results will be presented at the conference. This study pioneers the use of Rasch measurement theory to develop invariant measures for assessing sense of belonging in higher education, advancing research and practice in student development and institutional support.

Keywords: Sense of belonging, Rasch measurement theory, reliability, validity, fairness

Generating math word problems aligned with pupil ability and item difficulty

by Jie Wang | Central China Normal University, Wuhan, China

Xinguo Yu | Central China Normal University, Wuhan, China

Abstract ID: PROMS2025-CN006

Presenter Name(s): Xinguo Yu

Abstract:

This study proposes a knowledge-enhanced framework for the automatic generation of mathematical word problems, targeting elementary school students with the goal of producing problems that are diverse, logically coherent, and appropriately challenging for their cognitive level. Existing generation methods often lack alignment with instructional goals and struggle to control difficulty effectively, particularly in meeting the needs of personalized education at the primary level. To address this gap, the proposed framework integrates a structured core mathematics knowledge graph into a T5 pre-trained language model, ensuring that the generated problems adhere to curricular logic and pedagogical objectives. A coverage vector mechanism is introduced to dynamically track and regulate numerical content, thereby improving both mathematical consistency and problem diversity. Furthermore, domain-specific data augmentation techniques—such as synonym replacement and terminology transformation—are employed to enhance linguistic variation while preserving semantic precision and age-appropriate comprehension. For difficulty control, the study develops a structured five-dimensional evaluation model tailored to elementary-level word problems, encompassing contextual complexity, arithmetic level, reasoning depth, number of knowledge points, and word count. This model enables fine-grained difficulty assessment, supports the construction of difficulty-equivalent test papers, and can be integrated with the Rasch model to estimate student ability and design adaptive assessments. Once problem difficulty is quantified through this multi-dimensional framework, the Rasch model can reliably infer latent ability parameters from student response data. Experimental results on the GSM8K dataset demonstrate that the proposed method surpasses T5-small and T5-base in terms of accuracy and ROUGE scores, and approaches the generation quality of GPT-3.5. The minimal variance in difficulty across multiple dimensions further confirms the framework's effectiveness in maintaining consistent difficulty throughout batch generation. By integrating generation and difficulty evaluation into a unified, closed-loop system, the framework offers a scalable and pedagogically aligned solution for intelligent mathematics education.

Keywords: Math word problem generation, problem difficulty evaluation, knowledge enforced, coverage vector, intelligent education

Multiple intelligence assessment for rural primary students: Promoting equity through gamified-designed and non-graded inventory

by Kaixin Liang | Guangdong University of Foreign Studies
Wen Qin | Guangdong University of Foreign Studies
Rou Chen | Guangdong University of Foreign Studies
Ziqi Li | Guangdong University of Foreign Studies
Xiaomin Mai | Guangdong University of Foreign Studies

Abstract ID: PROMS2025-CN007

Presenter Name(s): Kaixin Liang

Abstract:

In China, the trend of 'not labelling' has emerged in the assessment of children's performance and competencies in primary schools. While it remains important to provide feedback to children and their educators based on competencies, this trend advocates for equality among children with diverse socioeconomic status (SES), characteristics, and abilities throughout the assessment process, aiming to provide constructive feedback to children and educators, highlighting strengths and fostering growth without attaching labels. In response to this trend, we aim to develop a non-labelling assessment tool, the Gamified Multiple Intelligence Inventory (GMII), rooted in Gardner's Multiple Intelligence Theory (MI) and based on gamified-designed tasks and procedures. Assessing seven intelligence domains – linguistic, logical-mathematical, spatial, musical, interpersonal, intrapersonal, and naturalist to allow for a holistic understanding of student competencies, this tool is to help educators identify a range of competencies without stigmatization and support personalized growth strategies. GMII comprises P-section, 30-minute paper-based tasks (e.g., scenario questions) in class, and T-section, 40-minute interactive/hands-on tasks using physical objects (e.g., building challenges) conducted in a standardized game room. 120 students (Grades 1–3; 40/grade) are randomly included from a rural primary school participating in International Collaboration for Integrated English Program in Guangdong. Students' responses are compiled from their answer sheets for the P-section and from video recordings for the T-section. To ensure reliability and validity, video recordings will be coded for behavioral analysis. Semi-structured interviews will be conducted with 60 students and all researchers to assess content validity. The assessment's reliability was confirmed through inter-rater reliability. Intraclass Correlation Coefficient (ICC) for coded behaviours and internal consistency (Cronbach's α) for task clusters. Content validity is established via expert review, and construct validity is verified through confirmatory factor analysis (CFA), aligning with MI. We anticipate that the responses from students will reveal nuanced competency distributions across intelligences. Results from this study provide initial support for the GMII as a tool for assessing children's multiple intelligences.

Keywords: Multiple intelligences theory, gamified evaluation, competency equity, rural education, early primary students

Assessing differential rater functioning with the many-facet latent space Rasch model

by Kuan-Yu Jin | Hong Kong Examinations and Assessment Authority

Abstract ID: PROMS2025-HK001

Presenter Name(s): Kuan-Yu Jin

Abstract:

Differential rater functioning (DRF) refers to situations where human raters systematically assign different rating scores to individuals based on factors unrelated to the actual performance or quality being assessed. These factors might include the rater's biases or preferences across different subgroups, such as gender, race, or other characteristics. DRF should matter to anyone who cares about fairness, accuracy, or reliability in evaluations. More advanced psychometric tools are desired to study this underexplored topic. To date, there are not many measurement models are available to quantify rater effects. The most famous of these is the many-facet Rasch model (MFRM) and its extensions. When studying DRF, these models focus on a binary grouping variable that can be explicit (Engelhard & Wind, 2018; Jin & Eckes, 2022) or implicit (Jin & Wang, 2017). However, such ratee-rater interactions could be more random than expected, and the effects may differ for ratees belonging to the same particular explicit or implicit group. As latent space item response models (e.g., Jeon et al., 2021) have been developed to account for item-person dependencies, in this study, the many-facet latent space Rasch model (MFLSRM), which assumes that raters would give different degrees of penalties in their ratings depending on the distance between the rater and the ratee, is proposed to quantify these intricate interactions. An experimental data in which raters were invited to evaluate the positive impression of facial photographs was selected to illustrate the utility of the new model. For this data, Bayesian model-data fit indices favored the proposed MFLSRM over the MFRM. The MFLSRM successfully yielded the relative positions between raters and ratees in the estimated latent space. In addition to the fact that raters may exhibit DRF according to the grouping variables, this study also revealed another point that raters may give biased ratings according to their own positions.

Keywords: Differential rater functioning, latent space, facets, Bayesian estimation

Unpacking student performance in visual arts: A many-facet Rasch analysis

by Joseph Chow | Hong Kong Examinations and Assessment Authority

Kuan-Yu Jin | Hong Kong Examinations and Assessment Authority

Abstract ID: PROMS2025-HK002

Presenter Name(s): Kuan-Yu Jin

Abstract:

This study explores the multifaceted influences on student performance in the 2023 Hong Kong Diploma of Secondary Education (HKDSE) visual arts test. It examines how factors such as student ability, assessment criteria, rater variability, language proficiency, and creative themes affect scoring outcomes. Grounded in the many-facet Rasch model (MFRM), this research extends Rasch measurement theory to account for multiple sources of variability in performance assessments. This framework allows for the disentangling of complex interactions among facets, providing insights into fairness, reliability, and validity in subjective evaluations, particularly in visual arts assessments. Data from the 2023 HKDSE visual arts examination, involving approximately 2,100 candidate works rated by trained examiners, was analyzed. The exam paper analyzed featured five questions, each allowing candidates to engage with reproduction artwork and utilize reference materials. The analysis focused on Part A, where candidates worked in two dimensions using any media, style, or technique, selecting one question to critically appreciate and analyze provided artworks. MFRM was employed to calibrate various factors influencing scores, including student abilities, assessment criteria, rater differences, language (English/Chinese), and creative themes. Findings indicated that the complexity of creative themes and rater severity significantly impacted scores, with certain themes being more challenging and specific raters consistently stricter. Language proficiency notably affected performance, especially for students assessed in their second language. Criteria related to technical skills exhibited less variability than those concerning creativity, with student ability identified as the strongest predictor of outcomes. Fit statistics demonstrated a good model-data fit, though minor inconsistencies among raters were noted. This study underscores the utility of MFRM in unpacking assessment dynamics and reveals biases linked to raters and themes. It suggests the necessity for enhanced rater training, theme standardization, and language support to foster equity in visual arts testing. These insights contribute to educational measurement by demonstrating how multifaceted analyses can provide a sophisticated understanding of high-stakes assessments.

Keywords: Visual arts assessment, many-facet Rasch model, student performance, rater variability, high-stakes assessments

Bridging theories and practice of inclusive assessment: A systematic review of frameworks and measures

by Jiaying Chen | The Education University of Hong Kong

Jiahe Gu | The Education University of Hong Kong

Zi Yan | The Education University of Hong Kong

Abstract ID: PROMS2025-HK003

Presenter Name(s): Jiaying Chen

Abstract:

Inclusive assessment, integral to inclusive education, serves multiple functions from enhancing learning outcomes to fostering a sense of belonging among diverse student groups. To effectively support and evaluate such practice, the development and implementation of precise and reliable measurement tools are crucial. These tools play a critical role in assessing the inclusion in education, thereby promoting students' engagement and academic success. Thus, this study focuses primarily on a systematic review of measurement tools for inclusive assessment, while also examining relevant frameworks to provide theoretical understandings. For the measurement tools, we particularly evaluate their psychometric evidence according to the Standards for Educational and Psychological Testing. Our literature search on ERIC, Web of Science, Scopus and PsycInfo followed the PRISMA guidelines, with a search string incorporated pertinent concepts such as differentiated assessment to ensure thorough coverage. Our preliminary results suggest that, measurement tools focusing specifically on inclusive assessment are limited, and demonstrate only certain degrees of validity (e.g., test content validity) and reliability evidence (e.g., internal consistency). The Classical Test Theory was the main approach for scale development. And in most cases, items about inclusive assessment are scattered within different measurements. Additionally, inclusion in assessment is frequently interpreted as a multi-dimensional concept, such as students' access and participation, from diverse discourses like the social political model of disability. Summarising findings at this stage, two major challenges to bridge inclusive assessment theories and practices are: (1) absence of affordable frameworks that could be translated into specific and measurable dimensions and items, and (2) lack of valid measures. To tackle these challenges and enhance the accessibility and implementability of inclusive assessment practice, we take a cultural and contextual standing to propose our inclusive assessment framework from the perspectives of Taoism, a prominent philosophy in Chinese society, and the current evolving learning environment, which characterised by advanced technologies such as Generative AI. Such framework could serve as a foundation for developing and validating a scale for inclusive assessment practice. We hope that, this study could shed light on inclusion around and in assessment, inform pedagogical practice, and guide future research directions.

Keywords: Inclusive assessment, measurement, assessment framework, review

Variable-length multistage adaptive testing design

by Chia-Ling Hsu | Hong Kong Examinations and Assessment Authority

Abstract ID: PROMS2025-HK004

Presenter Name(s): Chia-Ling Hsu

Abstract:

Multistage adaptive testing (MST) offers a balanced approach to adaptability, practicality, measurement accuracy, and control over test constraints. Consequently, MST has gained prominence in large-scale international assessments, such as the Programme for International Student Assessment (PISA), the Programme for the International Assessment of Adult Competencies (PIAAC), and the National Assessment of Educational Progress (NAEP). In MST, routing decisions to subsequent stages are primarily determined by estimating an examinee's responses within a given stage. Therefore, ensuring measurement precision in estimating an examinee's latent trait at each stage is critical for effective routing. Similar to item-level adaptive testing (commonly referred to as computerized adaptive testing; CAT), the precision of an examinee's latent trait estimate serves as an indicator of measurement accuracy in MST. This study conducted a series of simulation studies to compare various fixed-precision rules (also referred to as stopping rules) in terms of their effectiveness in recovering true ability estimates and optimizing test length in MST. Since examinees administered different test lengths to terminate MST upon achieving a pre-specified precision, variable-length MST is used to distinguish it from fixed-length MST. The stopping rules examined include maximum standard error (SE) rule (also known as the minimum information rule), absolute change in theta (CT) rule, minimum information rule, and joint rule. Additionally, the study manipulated factors such as the number of items available for test assembly, the maximum number of items administered, and the distribution of item characteristics across MST stages. The simulation results showed that a more stringent precision criterion enhanced measurement precision but reduced test efficiency. Specifically, the CT rule required a longer average test length than the SE rule to achieve comparable precision. However, the CT rule improved the test efficiency of the joint rule (i.e., SE-CT rule) when the SE rule is dominant. Furthermore, a less stringent precision criterion in earlier stages is sufficient when a strict criterion is applied in later stages. In sum, the simulation findings provide valuable insights into the utility of different stopping rules across various scenarios of variable-length MST.

Keywords: Multistage adaptive testing, standard error, fixed precision, stopping rule, variable-length

The Indonesian Food Neophobia Scale revisited: A two-year Rasch-based study and differences in family eating habits

by Itsar Bolo Rangka | Universitas Negeri Malang, Indonesia

Abstract ID: PROMS2025-IN001

Presenter Name(s): Itsar Bolo Rangka

Abstract:

After three decades since the introduction of the Food Neophobia Scale (FNS) by Pliner and Hobden (1992), the Indonesian version of the FNS (ID-FNS) was successfully validated for the first time in 2023 using the Rasch Model. This study aims to (1) re-evaluate the stability of the psychometric properties of the ID-FNS after two years of use and (2) examine differences in food neophobia levels based on family eating habits among a sample of Indonesian adults. Food neophobia has been associated with risks related to nutritional adequacy and metabolic risk factors due to limited food choices, picky eating behaviours, dietary outcomes, and parental feeding practices (Howard et al., 2012; Tuorila et al., 2001). This is a cross-sectional study involving 621 adults (Mage = 26.011, SD = 7.144) from 24 provinces across Indonesia. Participants completed the 10-item Indonesian Food Neophobia Scale (ID-FNS) online and reported their daily family meal practices. To assess the stability of the psychometric properties of the ID-FNS after two years of use, a linking process within the Rasch Model was applied by anchoring the previous 10-item ID-FNS (2023) to the latest dataset (2025). Additionally, an analysis of variance (ANOVA) was conducted to compare the food neophobia scores between families with traditional and modern eating habits. The results of the linking process indicate that there were no excessive changes in the characteristics of the 10-item ID-FNS administered to 1,632 participants in 2023 compared to 621 participants in 2025. The 10-item ID-FNS also demonstrated measurement invariance across gender and exhibited adequate construct validity. A significant difference was found in food neophobia scores between families adopting traditional eating patterns and those with modern eating habits (p < .05; Cohen's d = .20). Families with traditional eating habits exhibited higher mean food neophobia scores (-0.49 logit) compared to those with modern eating habits (-0.66 logit). Several technical notes are also included as key findings of this study. Our study confirms the two-year stability of the ID-FNS. Additionally, it addresses issues related to food neophobia, including dietary diversification, food acceptance, and nutrition in both traditional and modern families.

Keywords: Food Neophobia Scale, Indonesian version, Rasch linking process, test-retest study, family eating habits

Investigating potential differential item functioning on the Hating Adolescence Test (HAT) using Rasch model

by Nila Zaimatus Septiana | IAIN Kediri

Intan Nuyulis Naeni Puspitasari | IAIN Kediri

Choirul Annisa | IAIN Kediri

Rizqona Maharani | IAIN Kediri

Choiru Ummatin | IAIN Kediri

Ummiy Fauziyah Laili | IAIN Kediri

Agus Miftakus Surur | IAIN Kediri

Suharni | Universitas PGRI Madiun

Abstract ID: PROMS2025-IN003

Presenter Name(s): Nila Zaimatus Septiana

Abstract:

The Hating Adolescents Test (HAT) is a concise self-report instrument developed to assess hatred among adolescents. Although it exhibits satisfactory psychometric properties, it is imperative to investigate potential measurement bias across diverse cultural contexts, such as within Indonesia. The analysis of differential item functioning (DIF) in this context is crucial for ensuring equitable measurement across various subgroups at the item level, representing a fundamental aspect of construct validity. This study examined differential item functioning (DIF) within the Indonesian adaptation of the Hating Adolescents Test (HAT) concerning gender, age, and residential location, employing the Rasch model. The aim was to ensure equitable measurement of hatred among diverse adolescent subgroups in Indonesia. Questionnaire data were collected from a total of 1,325 senior high school students (aged 13-18 years) who were randomly sampled from various urban and rural schools with permission from the school authorities and informed consent from the participants and their parents/guardians. Rasch analysis was utilized to identify DIF by comparing item difficulty parameters across the defined subgroups (gender, age, and residence) using WINSTEPS software. The magnitude of DIF was assessed by DIF contrast and statistical significance. Findings from the gender-based analysis revealed a statistically significant bias in Item 11 (DIF contrast = 0.72, p < 0.01), indicating a higher propensity among male respondents to endorse negative statements concerning body weight. Furthermore, the age-based analysis demonstrated substantial DIF in Item 1 for the 17-year-old subgroup (DIF contrast = 0.69). Conversely, no evidence of DIF was observed based on the participants' place of residence. These findings underscore the salient influence of gender and age on the perception of specific items, thereby necessitating careful adjustments in the development of assessment instruments to ensure fairness and accuracy of measurement across diverse populations. Future research should explore additional factors contributing to response discrepancies and meticulously consider socio-cultural contexts during the design phase of measurement tools.

Keywords: Differential item functioning (DIF), Rasch model, item bias, hatred, adolescents

The expert judgement validation of Student Growth Mindset Scale (SGMS) using Many Facet-Rasch Measurement (MFRM)

by Ma'rifatin Indah Kholili | Universitas Pendidikan Indonesia, Bandung, Indonesia

Nandang Rusmana | Universitas Pendidikan Indonesia, Bandung, Indonesia

Ahman | Universitas Pendidikan Indonesia, Bandung, Indonesia

Nandang Budiman | Universitas Pendidikan Indonesia, Bandung, Indonesia

Rahmi Ramadhani | Universitas Potensi Utama, Medan, Indonesia

Abstract ID: PROMS2025-IN005

Presenter Name(s): Ma'rifatin Indah Kholili

Abstract:

This study aims to validate the Student Growth Mindset Scale (SGMS) through expert assessment. This study explores the statistical overview of the MFRM analysis, rater measurements related to dimensions and criteria, rater severity and leniency in assessing Scale quality. This study applies the principle of psychometric content and the reliability of the assessor validation, to determine the validity and reliability of SGMS for assessing students' growth mindsets. MFRM was chosen because it can accommodate assessment variability caused by many raters. This study used Many-Facet Rasch Measurement (MFRM) to analyze the Growth Mindset Scale assessment. The data analyzed involved 12 dimensions-aspects, 3 assessment criteria (usability, feasibility, and accuracy), and 7 raters. The instrument used in this study is the rubric of dimension and aspect assessment on the Growth Mindset Scale. The validation process involved seven experts. The analysis was conducted with FACETS software version 3.84.0 and included adjusting the mathematical model to include interaction effects based on the rater's gender and scientific background. The results of the expert assessment show that the Growth Mindset Scale has several weaknesses related to the variation in the quality of the dimensions and the presence of rater bias. There is a bias in the assessment based on the rater's gender and academic background. Male raters tend to be stricter in judgment than female raters, and raters with a psychology background tend to be different in judgment than raters with counselling guidance backgrounds. The "Usability" criterion is rated as the most challenging criterion to apply by the rater. Overall, the results of the expert assessment show that the Growth Mindset Scale has several drawbacks related to the variation in the quality of the dimensions and the presence of rater bias. Recommendations for follow-up research are to expand the number and variety of rater backgrounds; integrate qualitative approaches, such as interviews or think-aloud protocols, to dig deeper into the cognitive and affective processes behind rater assessment; and apply multifaceted models in cross-cultural contexts to test the consistency of rater bias.

Keywords: Growth mindset, Many Facet-Rasch Measurement, student, expert validation

Evaluating students' performance on cryptarithms: Item analysis from a pilot study

by Elizar | Universitas Syiah Kuala

Anwar | Universitas Syiah Kuala

Ayu Mastura | Universitas Syiah Kuala

Abstract ID: PROMS2025-IN008

Presenter Name(s): Elizar

Abstract:

This pilot study explores senior high school students' problem-solving skills through cryptarithm problems, mathematical puzzles where digits are replaced by letters or symbols, requiring solvers to determine the correct numerical values. Cryptarithms enhance logical reasoning, pattern recognition, and creative thinking, making them valuable tools for assessing higher-order cognitive skills in mathematics education. The study aimed to analyze the quality of three cryptarithm items and evaluate students' performance using the Heuristic Problem Solving framework. Data were collected from 30 senior high school students, each completing three cryptarithm problems. The data were analyzed using Jmetrik. Item difficulty, discrimination, and test reliability were examined to determine the validity of the problems. Findings revealed that overall student performance was low, indicating a need for improved instructional support in problem-solving tasks. However, the items exhibited acceptable psychometric characteristics, suggesting they are suitable for inclusion in further study. As a pilot study, this research provides initial insights into students' challenges with cryptarithms. The finding will be used for the need analysis to justify the need for developing online learning materials to promote students' problem solving skills in cryptarithm. This study supports the integration of innovative mathematical tasks to cultivate critical thinking, logic, and creativity among high school learners.

Keywords: Cryptarithm, problem-solving, item analysis

Teachers' perception about nature of science: A Rasch model measurement analysis

by Kartimi | UIN Siber Syekh Nurjati, Cirebon, Indonesia

Siti Nadya Zynuddin | Universiti Malaya

Bambang Sumintono | Universitas Islam Internasional Indonesia

Abstract ID: PROMS2025-IN010

Presenter Name(s): Kartimi, Siti Nadya Zynuddin, Bambang Sumintono

Abstract:

Globally, teachers' understanding of the Nature of Science (NOS) is vital for enhancing scientific literacy. This is because teachers' perspectives on NOS significantly shape their approach to science instruction and can profoundly influence students' comprehension and success in the subject. In Indonesia, student achievement in science remains a concern. National public examination results in science subjects consistently lag behind other subjects, and international comparative studies such as TIMSS and PISA have shown that Indonesia's rankings have not improved as expected. This study aims to explore Indonesian science teachers' perceptions of NOS. It employs a cross-sectional, non-experimental design with a quantitative approach, focusing on objective measurement using the Rasch Model for data analysis. The primary instrument used is the Reconceptualized Family Resemblance Approach to Nature of Science Questionnaire (RFNQ) developed by Erduran, which comprises eleven constructs related to the understanding of NOS and consists of 70 items measured using a four-point Likert scale.Demographic data—including gender, age, educational background, and subject taught—were also collected. The questionnaire was distributed electronically via an online platform (Google Forms). Data collection is still ongoing. At this stage, responses have been obtained from 200 science teachers, primarily from the West Java province of Indonesia. Preliminary results indicate that the data exhibit acceptable reliability indices, good construct validity, and a functional rating scale. Notably, item analysis reveals varying difficulty levels across constructs, offering insights into Indonesian teachers' perceptions of science. Inferential statistical tests also highlight differences across demographic variables.

Keywords: Nature of science, Rasch model, person analysis, Indonesian school, science education

Development of a web-based ESQ assessment tool using Rasch model analysis for holistic psychological well-being

by Basma Tania | Universitas Negeri Malang, Indonesia
Kukuh Setyo Pambudi | Universitas Negeri Malang, Indonesia
Jati Fatmawiyati | Universitas Negeri Malang, Indonesia
Iffat Maimunah | UIN Maulana Malik Ibrahim Malang
Wildana Wargadinata | UIN Maulana Malik Ibrahim Malang
Tutut Chusniyah | Universitas Negeri Malang, Indonesia
Mochammad Said | Universitas Negeri Malang, Indonesia
Muhammad Izzudin Haq | Center for Social Psychology and Society
Syabiilah Azzahroh Widyatmoko Putri | Universitas Negeri Malang, Indonesia
Habil Abyad | Universitas PTIQ Jakarta

Abstract ID: PROMS2025-IN013

Presenter Name(s): Basma Tania, Kukuh Setyo Pambudi

Abstract:

Emotional and Spiritual Quotient (ESQ) is conceptualized as an integrated psychological construct encompassing emotional and spiritual capacities that contribute to holistic well-being. Recognizing the importance of these dimensions in supporting mental health, this study aims to develop a valid and reliable web-based assessment tool for measuring ESQ comprehensively. A sequential mixed-methods approach is employed, beginning with a qualitative bottom-up phase that involves collecting and refining ESQ concepts through expert consultations and an extensive literature review. This foundational phase ensures the content validity and contextual relevance of the instrument. Subsequently, two pilot studies are planned, each involving 300 participants. The first pilot study focuses on item calibration and domain bias assessment to identify and adjust problematic items. The second pilot study aims to refine the instrument's psychometric properties, including validity, reliability, and item discrimination, with validity specifically analyzed using the Rasch Model. This rigorous psychometric evaluation is intended to produce a measurement tool with strong construct validity and reliability. In parallel, a user-friendly web-based platform will be developed to facilitate broad accessibility and efficient data collection. The platform is designed to provide an interactive and seamless experience for respondents, supporting robust data management and scalability. The expected outcome of this research is a psychometrically sound ESQ assessment instrument that can be applied in clinical, educational, and personal development contexts to enhance holistic psychological well-being. Moreover, the tool is anticipated to enable future research exploring the relationships between ESQ and other psychological and health-related variables, thereby contributing to a deeper understanding of holistic well-being.

Keywords: Emotional and spiritual quotient, web-based assessment, psychological well-being

Development and validation of a Perceived Practical Teaching Competence Scale (PTCS) for middle school students in science classes using the partial credit model and the confirmatory factor analysis

by Sun-geun Baek | Seoul National University
Woori Song | Seoul National University
Yunah Kang | Seoul National University
Byunghoon Jeon | Seoul National University
Seojin Kim | Seoul National University

Abstract ID: PROMS2025-KR001

Presenter Name(s): Sun-geun Baek, Woori Song

Abstract:

This study aims to develop and validate a practical teaching competence scale(PTCS) for middle school students in science classes using the partial credit model(PCM) and the confirmative factor analysis(CFA). Practical teaching competence refers to the ability to perform effectively in real teaching situations to successfully carry out subject instruction. Based on a comprehensive literature review, five sub-domains were newly set as 'Planning and Organization', 'Communication', 'Interaction', 'Coordination', 'Sincerity and Enthusiasm', and a preliminary test consisting 30 items (6 items for each sub-domain) was developed. In addition, 10 PhDs in educational measurement and evaluation reviewed the preliminary test, and each item achieved a content validity index (CVI) exceeding 4.50 out of 5.00. A preliminary test was conducted on 266 middle school students and construct validity and reliability were checked to complete a scale consisting of 15 items. To this end, four competitive models were made: (1) 15 items considering only the item fit, (2) 15 items considering the item fit within sub-element, (3) 15 items considering the item fit within sub-domain, (4) 15 items considering the item fit within sub-domain and sub-element. Construct validity and reliability were compared. As a result, the third model showed the best construct validity (model fit: RMSEA = 0.058, TLI = 0.968, CFI = 0.974) and a good reliability (Cronbach's α= 0.955). This study developed and validated a scale to reliably and validly measure students' perceptions of their teacher's practical teaching competence. This study is significant in that it provides a foundation for systematically understanding students' perceptions of their teacher's teaching competence in the context of science classes and for diagnosing and improving practical teaching competence. This study is significant in that it provides a foundation for systematically understanding students' perceptions of their teacher's teaching competence in the context of science classes. Furthermore, the developed scale can serve as a useful tool for identifying specific areas where teachers may need support, guiding their professional growth through targeted training, and fostering continuous improvement in their instructional practices.

Keywords: PCM, CFA, practical teaching competence scale, content validity, construct validity

Evaluating the fairness of a high-stakes college entrance exam in Kuwait: A Rasch model application

by Fajer Shamsaldeen | Kuwait University

Jue Wang | University of Science and Technology of China

Soyeon Ahn | University of Miami

Abstract ID: PROMS2025-KU001

Presenter Name(s): Jue Wang

Abstract:

The use of college entrance exams for facilitating admission decisions become controversial, and the central argument is around the fairness of test scores. The Kuwait University English Aptitude Test (KUEAT) is a high-stakes test, but very few studies have examined the psychometric quality of the scores for this national-level assessment. This study illustrates how measurement approaches can be used to examine the fairness issues in educational testing. Through a modern view of fairness, we first calibrate the KUEAT items and obtain latent scores for individual students based on Rasch measurement theory. We then assess the internal and external bias of KUEAT scores specifically using differential item functioning analysis and differential prediction analysis and provide a comprehensive fairness argument for KUEAT scores. The analysis for examining the internal evidence of bias was based on 1790 examinees' KUEAT scores in November 2018. KUEAT scores and first-year college GPAs of 4033 students enrolled in KU were used for assessing the external evidence of bias. Results revealed many items showing differential item functioning across student subpopulation groups (i.e., nationality, gender, high school majors, and high school types). Meanwhile, KUEAT scores also predicted college performance differentially by different student subgroups (i.e., nationality, high school majors, and high school types). Discussion and implications on the fairness issues of college entrance tests in Kuwait are provided.

Keywords: High-stakes college entrance exam, fairness, differential item functioning, Rasch measurement theory, differential prediction analysis

Validation of a burnout assessment tool for healthcare workers: A psychometric approach using Rasch modelling and exploratory factor analysis

by Suriya Kumareswaran Vallasamy | National University of Malaysia

Rosnah Ismail | National University of Malaysia

Abstract ID: PROMS2025-MY002

Presenter Name(s): Suriya Kumareswaran Vallasamy

Abstract:

Burnout is a major and growing concern among healthcare workers (HCWs), exacerbated by high workloads, emotional stress, and cognitive fatigue. Given the demanding nature of healthcare environments, early detection and intervention are essential. This study aimed to validate a newly developed burnout assessment tool specifically tailored for HCWs, focusing on two core dimensions: cognitive exhaustion and emotional exhaustion. The theoretical foundation of the scale draws upon the dual-process model of occupational burnout, which posits distinct yet interrelated emotional and cognitive components. A cross-sectional survey was conducted among 254 HCWs in Malaysia, specifically from the Johor Bahru District Health Office and Hospital Sultanah Aminah. The 20-item scale underwent Exploratory Factor Analysis (EFA) using Varimax rotation, which revealed a stable two-factor structure with item loadings ranging from 0.42 to 0.81. Internal consistency was excellent, with Cronbach's alpha values between 0.95 and 0.97. Rasch analysis complemented the EFA findings, providing strong evidence of unidimensionality within each subscale and acceptable item fit (infit/outfit MNSQ = 0.5–1.5), along with high point-measure correlations. Misfitting responses (Entries 2, 22, and 33) were identified and removed to improve model precision. Both subscales demonstrated excellent reliability, with person separation indices above 3.6 and item separation above 4.3, alongside well-ordered response categories. Response category probability curves confirmed distinct thresholds across the five-point Likert scale. Differential Item Functioning (DIF) analysis indicated minor gender- and race-related item biases; however, item performance remained within acceptable psychometric thresholds. Person-item maps illustrated a well-aligned distribution of item difficulty and respondent ability. The explained variance exceeded 69% in both subscales, with minimal residual variance, confirming strong measurement unidimensionality. These findings affirm that the burnout scale is a psychometrically sound instrument for assessing cognitive and emotional exhaustion in HCWs, supporting its use for preventive screening and targeted mental health interventions.

Keywords: Burnout, healthcare workers, Rasch model, exploratory factor analysis, psychometric validation

Automated PPKI application system via Google forms and WhatsApp: Development and evaluation using Rasch measurement model

by Mei Teng Ling | Miri District Education Office

Nur Hanini Anne Abdullah | Miri District Education Office

Felicia Suling Emang | Miri District Education Office

Abstract ID: PROMS2025-MY003

Presenter Name(s): Mei Teng Ling

Abstract:

This study aimed to develop and evaluate an automated system for managing transfer and admission applications for Special Education Integration Program (PPKI) students. The objectives were to: (a) enable officers to track applications efficiently, (b) minimize missed or delayed applications, (c) allow parents to check their application status independently, (d) improve communication via automated WhatsApp notifications, and (e) assess officers' competency in resolving placement issues using the system. The study was guided by the Technology Acceptance Model (TAM) and public service delivery frameworks. The Rasch Measurement Model was used to validate binary (Yes/No) Guttman-scaled instruments for system evaluation. A developmental design was employed. The system was built using free platforms: Google Forms, Sheets, Apps Script, CallMeBot API, and Looker Studio. Three Guttman-type instruments were used: (a) a Parent Interaction Checklist to assess user experience, (b) a System Performance Observation Checklist to document automation outcomes, and (c) a 15-item Officer Competency Scale focusing on resolving issues such as incomplete applications, follow-ups, and school placement decisions. Data from 30 parents and 15 education officers were analyzed using the Rasch dichotomous model to assess reliability, item fit, and response patterns. The officer competency scale showed strong internal consistency (item reliability = 0.88; person reliability = 0.83). Most items showed good fit (MNSQ 0.75–1.28), indicating that the scale effectively measured a single latent trait—practical competency in using the system. Parent checklist results showed that 87% completed applications independently, and 93% received timely WhatsApp updates. Officers reported improvements in workflow and efficiency. The system successfully streamlined PPKI application handling and supported effective communication with parents. The Guttman-scaled instruments validated through Rasch analysis proved reliable in evaluating officer competency and system impact. The approach is cost-effective and scalable for broader implementation in education administration.

Keywords: PPKI application, Rasch measurement, officer competency, system automation, Google apps script

Expert validation of the C-A-RE module for sandwich generation workers using many-facet Rasch analysis

by Rozita Jayus | Universiti Teknologi Malaysia & University Malaysia Terengganu

Aqeel Khan | Universiti Teknologi Malaysia

Adibah Abdul Latiff | Universiti Teknologi Malaysia

Mastura Mahfar | Universiti Teknologi Malaysia

Nornazira Binti Suhairom | Universiti Teknologi Malaysia

Siti Aminah | Universitas Negeri Yogyakarta, Indonesia

Abstract ID: PROMS2025-MY005

Presenter Name(s): Rozita Jayus

Abstract:

This study examined the psychometric validity of the C-A-RE module, a specific programme to reduce stress and increase career adaptability and career resilience in sandwich generation workers at public universities who are simultaneously caring for ageing parents and dependent children and managing job responsibilities. Eight experts were assessed using the Many-Facet Rasch Model (MFRM). Five of the eight experts were lecturers in counselling and psychology, and three of them were counsellors with many years of experience. The experts assessed six key items covering content relevance, feasibility and expected impact on the target group. The MFRM analysis showed robust psychometric properties, with the variance explained by the Rasch measures reaching 52.06% and the observed exact agreement of 75% being very close to the expected agreement of 76.1%. The reliability indices showed moderate to high consistency with values of 0.60-0.65 for the raters and 0.68-0.82 for the items in the population and sample calculations. The chi-square of the fixed model for the items $X^2=6.6$, df=5, p=0.25 indicates a good fit of the model. The analysis at item level revealed particularly strong performance in the dimensions of stress, career adaptability and career resilience with exceptionally high point-measure correlations of 0.92, 0.88 and 0.92 respectively and an exact agreement of 80% in each case. The alignment of the module with the needs of the target group received a high level of agreement (77.5%), while the appropriateness of the time allocation led to a comparatively lower level of agreement (57.5%). The experts' assessments were predominantly positive. Three experts rated the module as high (extreme) and three as high, resulting in an average overall rating of 4.67 out of 5. Only two experts gave moderate ratings. Overall, these results confirm that the C-A-RE module has good expert validity, which emphasises its suitability for implementation to support the wellbeing and career development of sandwich generation workers in higher education.

Keywords: Experts validation, module development, Many-Facet Rasch Model (MFRM), C-A-RE module

Score linking and validation in educational tests: A Rasch model study

by Zouh Fong Chieng | Ministry of Education Malaysia

Abstract ID: PROMS2025-MY006

Presenter Name(s): Zouh Fong Chieng

Abstract:

This study aims to validate the interpretation of linking scores within the framework of the Rasch model. Four test forms were developed, each consisting of 45 multiple-choice items, including 8 anchor items shared across forms to facilitate linking. The instruments were administered to a sample of 449 students. Rasch analysis was conducted to calibrate item difficulties and estimate person abilities on a common scale. Model fit was evaluated using item fit statistics, with over 95% of items falling within acceptable thresholds. The standard error of measurement was examined across the ability continuum to support fine-grained score interpretations beyond the overall construct level, providing evidence for the precision of measurement at different points along the scale. Linking scores were interpreted as invariant representations of student ability, consistent with the Rasch model's principle of specific objectivity. These scores enabled meaningful inferences about expected performance patterns across test forms. In addition, item difficulty was examined in relation to item features, supporting construct-relevant interpretations and the structural validity of the instruments. This study employed a Nonequivalent Groups with Anchor Test (NEAT) design. Group A received Form S1 and Anchor Z, Group B received Form S2 and Anchor Z, Group C received Form S3 and Anchor Z, and Group D received Form S4 and Anchor Z. Since the groups were not randomly equivalent but all completed the same anchor test, linear equating was applied to place scores from Forms S1, S2, S3, and S4 onto a common scale. The results confirmed the psychometric equivalence of the four forms, the stability of item parameter estimates, and the effectiveness of anchor items in supporting score linking. This study contributes to educational measurement by demonstrating how Rasch-based linking can support valid and theory-driven interpretations of test scores across multiple forms, enhancing fairness, comparability, and interpretability in large-scale assessments.

Keywords: Rasch model, score linking, educational assessment, test validation, item response theory

Psychometric validation of a 21st century skills instrument in a design thinking context among final year polytechnic students using the Rasch measurement model

by Aede Hatib Musta'amal | Universiti Teknologi Malaysia

Nor Aisyah Che Derasid | Universiti Teknologi Malaysia

Mohd Safarin Nordin | Universiti Teknologi Malaysia

Nornazira Suhairom | Universiti Teknologi Malaysia

Rozita Jayus | Universiti Malaysia Terengganu

Abstract ID: PROMS2025-MY007

Presenter Name(s): Nornazira Suhairom

Abstract:

The increasing emphasis on 21st Century Skills in higher education has positioned Technical and Vocational Education and Training (TVET) as a critical pathway for equipping students with the competencies needed in today's innovation-driven workforce. However, despite policy support, evidence suggests that Malaysian polytechnic graduates often lack essential soft skills, such as critical thinking, collaboration, communication, creativity, and ethical decision-making. Design Thinking (DT), a user-centered and problem-solving methodology, has emerged as a promising framework to address this gap. This study reports on the psychometric validation of a newly developed instrument designed to measure 21st Century Skills among final-year polytechnic students engaged in DT-based final-year projects. Utilizing the Rasch Measurement Model, data from 211 students across two polytechnics were analyzed. The results showed that the instrument explained 42.0% of the total raw variance, with person and item reliability indices of 0.94 and 0.87, respectively. The scale demonstrated high internal consistency, as indicated by a Cronbach's alpha (KR-20) of 0.99, and robust item-person separation indices of 3.94 (person) and 2.60 (item). The unexplained variance in the first contrast was 6.1%, supporting the unidimensionality of the scale. Fit statistics for all items were within acceptable ranges, confirming the instrument's construct validity. These findings provide strong empirical support for the integration of Design Thinking into polytechnic curricula and offer a reliable tool for assessing skill development in TVET contexts. The study contributes to the advancement of assessment practices in education for the 21st century.

Keywords: Psychometric validation, design thinking, polytechnic student, 21st century skills instrument

The validation of integrating Artificial Intelligence construct for the multimodal learning framework development: A Rasch model measurement analysis

by Nurin Erdiani Mhd Fadzil | Universiti Kebangsaan Malaysia & Universiti Putra Malaysia

Harwati Hashim | Universiti Kebangsaan Malaysia & Eduxcellence: Development of Innovative Curriculum & Pedagogy Research Group

Abstract ID: PROMS2025-MY008

Presenter Name: Nurin Erdiani Mhd Fadzil

Abstract:

This study utilised the Rasch Model Measurement to evaluate the reliability and validity of the Artificial Intelligence construct within the proposed Multimodal Learning Framework. The pilot study involved 40 ESL foundation students and 283 ESL foundation for the actual study, focusing on assessing the measurement precision of the artificial intelligence-related items in the survey instrument. The Rasch model analysis, conducted using Winsteps software, examined item reliability, person reliability, and separation index to determine construct validity. The findings revealed an item reliability of 0.81 with a separation index of 1.92, indicating good internal consistency. The person reliability was recorded at 0.89 with a separation index of 2.85, demonstrating the instrument's ability to differentiate participant proficiency levels in adopting artificial intelligence-enhanced learning. The unidimensionality assessment confirmed the construct's validity, with unexplained variance within the acceptable range. Item fit analysis, including Mean Square (MNSQ) fit statistics and Point Measure Correlation (PMC), identified five items with misfit behaviour, which were revised to improve alignment with the construct's theoretical framework. These results affirm that the Artificial Intelligence construct is a valid and reliable measure for assessing students' engagement with AI in multimodal learning environments. Future studies should explore further refinements based on the findings to enhance measurement precision before large-scale implementation.

Keywords: Rasch model analysis, artificial intelligence, multimodal learning, construct validity, AI-enhanced learning

Path analysis of critical thinking in chemistry informed by the Rasch measurement framework

by Jonathan Barcelo | Saint Louis University

Abstract ID: PROMS2025-PH001

Presenter Name(s): Jonathan Barcelo

Abstract:

Chemistry concepts are fundamental to understanding many areas of health science professions. Hence, it is necessary to assess the critical thinking of health science students in their chemistry courses as chemistry-specific critical thinking impacts how students apply clinical reasoning in health-related contexts. Though many variables have been identified as important predictors of critical thinking in chemistry, the interrelationship of gender, competencies such as knowledge of chemistry concepts, knowledge of visual representations, ability to differentiate substances, and critical thinking in chemistry among health science students remain poorly explored. This study aimed to develop a model to describe the structural relationship of the abovementioned variables to health science students' critical thinking in chemistry. Anchored on the Heuristic-Analytic Theory of Reasoning, we hypothesized that critical thinking chemistry is influenced by the abovementioned variables as chemistry involves three domains of chemistry: macroscopic, submicroscopic, and symbolic. Furthermore, we also hypothesized that gender influences competencies in chemistry. Data was drawn from 577 second-year health science students in Baguio City, Philippines, who consented to participate in the study. These students have completed general chemistry, general organic chemistry, and analytical chemistry before data gathering. Using the Rasch analysis in Winsteps 4.4.5, we evaluated the unidimensionality, item fit, and differential item functioning of four research instruments: Prior Knowledge of Chemistry Concepts Test, Visual Representations Test, Chemical Identity Thinking Instrument, and Critical Thinking Test in Chemistry. Next, the student ability estimates (logits) were generated and exported to generate the path model in IBM SPSS Amos software. The path analysis revealed weak to moderate connections between the variables, although the model explained 31.0% of the variance in critical thinking in chemistry. The strongest predictor of critical thinking in chemistry was chemical identity thinking, followed by prior knowledge of chemistry concepts and then knowledge of visual representations. However, the strongest predictor of prior knowledge of chemistry concepts was knowledge of visual representations. The results suggest that improving health science students' chemical identity thinking can promote greater critical thinking in chemistry. Furthermore, it is also encouraged to include various visual representations when teaching chemistry concepts.

Keywords: Chemical identity, chemistry, critical thinking, health science, visual representations

Development and validation of the Visual Representations Test using Rasch measurement framework

by Jonathan Barcelo | Saint Louis University

Precious Lady Gine Araneta | Saint Louis University

Abstract ID: PROMS2025-PH002

Presenter Name(s): Precious Lady Gine Araneta

Abstract:

Visual representations in chemistry are important in promoting adequate understanding of chemistry concepts. Previous studies have emphasized their importance in understanding molecular phenomena, improving problem-solving skills and content knowledge, and sustaining students' attention and motivation. However, most non-chemistry undergraduate students in the Philippines have a poor understanding of visual representations in chemistry, necessitating the development of a research instrument that can measure representational knowledge. This study developed the Visual Representations Test (VRT), a research instrument designed to measure health science students' conceptual understanding of visual representations in chemistry. The construct, which is a conceptual understanding of visual representations, was hypothesized to progress in three levels: Level 1 (low conceptual understanding), Level 2 (marginal conceptual understanding), and Level 3 (adequate conceptual understanding). After securing ethics approval, a total of 404 first year to second year undergraduate students from four tertiary institutions in the Philippines were recruited during pilot testing. Second year undergraduate students have background in general and organic chemistry while first year students have not been exposed to any chemistry coursework. Dichotomous Rasch analysis was performed using WINSTEPS 4.4.5. to determine the construct validity and reliability of the research instrument. The results revealed satisfactory reliability (item reliability = 0.95, person reliability = 0.75) and separation (person separation = 1.74, item separation = 4.55). The results also provided evidence of item fit, construct validity, and unidimensionality. However, three items displayed differential item functioning between students with and without prior coursework in general and organic chemistry, suggesting areas for refinement of the items. Overall, the results suggest that the Visual Representations Test has the potential to serve as a diagnostic tool in non-chemistry undergraduate programs and inform the scientific community of learning activities that facilitate the progression of conceptual understanding of visual representations in general chemistry and organic chemistry.

Keywords: Chemistry, conceptual understanding, dichotomous Rasch analysis, visual representations

Development and Rasch analysis of the critical thinking test in chemistry

by Jonathan Barcelo | Saint Louis University

Mark Alben Ponciano | Saint Louis University

Abstract ID: PROMS2025-PH003

Presenter Name(s): Mark Alben Ponciano

Abstract:

Improving the critical thinking of undergraduate students in chemistry contexts is one of the desired competencies among health science programs. Critical thinking in chemistry courses relies on mechanistic reasoning, which is necessary to understand structure-property and structure-function relationships. However, an invariant measurement of critical thinking in chemistry is required to guide the design of appropriate teaching interventions in Philippine classroom settings. In this study, the Critical Thinking Test in Chemistry was developed and validated using a Rasch analysis. Drawing from the Heuristic-Analytic Theory of Reasoning, critical thinking in chemistry was conceptualized as the ability to link correct inferences and correct chemistry explanations in general inorganic chemistry and general organic chemistry contexts. The first version of the test was composed of 32 items involving concepts in basic general and organic chemistry but was reduced to 20 items during the final revision to reduce test anxiety and test fatigue, based on the recommendation of the evaluators. In each item, students were required to evaluate whether the conclusion in each statement is valid, possible, or invalid, then support their answer using chemistry-based explanations. A hypothesized progression of critical thinking in chemistry was drafted based on the theoretical construct. Students' answers were evaluated by two raters, then assigned a final score based on a rubric tool that evaluates the accuracy of answers, accuracy of explanations, and type of chemistry-based explanations used. After obtaining informed consent, the research instrument was administered to 969 first-year to second-year college students from 5 tertiary institutions in the Philippines. Then, Rasch analysis was performed using Winsteps 5.9.0 using the rating scale model. The results provided evidence of adequate construct validity, reliability, unidimensionality, local independence, and adequate item fit. However, there is a need to reduce the number of items in the research instrument based on the person-item map. The results of the analysis indicate that the Critical Thinking Test in Chemistry can be used as a diagnostic tool to measure the critical thinking of undergraduate students in chemistry within the Philippine setting.

Keywords: Chemistry, critical thinking, Rasch analysis, rating scale model

Application of Rasch analysis in the evaluation of biochemistry examination for health science students

by Jonathan Barcelo | Saint Louis University

Lloyd Allen Lorente | Saint Louis University

Abstract ID: PROMS2025-PH004

Presenter Name(s): Lloyd Allen Lorente

Abstract:

Biochemistry is a common pre-requisite course in health science programs across the Philippines, with term exams playing a crucial role in assessing the preparedness of students to study professional courses. However, ensuring that these exams are of high quality is essential for accurately assessing student knowledge of biochemistry concepts. This study evaluated the quality of 100 instructor-made multiple-choice items in a biochemistry exam using the responses of 203 health science students enrolled in one second-year undergraduate biochemistry course in February 2023. The questions were based on the table of specifications to determine the type of cognitive level per biochemistry topic. While item reliability and separation were adequate (item separation = 5.05; item reliability = 0.96), the person reliability and separation were low (person separation = 1.45, person reliability = 0.68). Four items also exhibited item misfit while seven items exhibited negative point measure correlation values. The unexplained variance in the first contrast was determined to be 3.29, indicating multidimensionality of the construct. The person-item map revealed that the item measures are within -4.16 to 6.74 while person measures were within -1.23 to 1.58, indicating that some items are too difficult for the students. The test was revised based on the recommendations of faculty members. After revising and deleting some items, the number of items was reduced to 77. The revised biochemistry test was administered to 243 second-year health science students last February 2024. Based on the results, the revised test had adequate person and item separation (person separation = 2.27; item separation = 5.18) and reliability (person reliability = 84; item reliability = 0.96). In addition, all items had adequate item fit and point measure correlation. While the eigenvalue of the first contrast was noted to be 2.35, the largest standardized residual correlation was only 0.24. Our findings demonstrate that Rasch analysis is a valuable tool for enhancing the quality and reliability of instructor-made biochemistry exams. We recommend that subject matter experts carefully review items before reusing them in future exams or depositing them into item banks.

Keywords: Biochemistry, Rasch analysis, reliability, validity

Game leveling using the Rasch model

by Tzemin Chung | CommonTown Pte Ltd

Mohd Zali Mohd Nor | Persatuan Rasch Malaysia

Richard Yan | CommonTown Pte Ltd

Peing Ling Loo | CommonTown Pte Ltd

Abstract ID: PROMS2025-SG001

Presenter Name(s): Tzemin Chung

Abstract:

This study aimed to develop a measurement scale for vocabulary game challenges to support deliberate practice for English second-language learners, ensuring challenges align with students' abilities for effective learning. Drawing on Self-Determination Theory, which highlights competence, autonomy, and relatedness as key drivers of engagement, this study posits that games effectively sustain student interest in learning. This guided the inclusion of games as part of deliberate practice within an ebook platform for vocabulary instruction, prompting the development of a measurement scale to ensure game challenges fit student abilities. The study asked: How can the Rasch model help to create an effective scale for game challenges? Five vocabulary games— The Wall (word recognition), Quick Speak (sentence reading), and Word Safari, Word Finder, and Word Catacombs (all focused on word generation from letters)—were designed, categorized into three CEFR difficulty levels (A1, A2, B1; Primary 1 to 6 and Secondary 1, or Grades 1 to 7). The study involved 700 students aged 5–13 from Singapore and Turkey, attempting 633 of 1465 challenges, yielding 5547 data points. After filtering for challenges with over 10 attempts, the Rasch partial credit model analyzed data from 668 students and 160 challenges using Winsteps. Rasch analysis showed a robust item hierarchy (item separation 5.36, reliability 0.97), with challenge difficulties from -4 to +8 logits. However, person separation (0.53) and reliability (0.22) were low, indicating limited differentiation of student abilities due to sparse data (averaging 8 responses per student) and low performance variation. The Wright Map revealed a narrow student ability distribution (-2 to +1 logits), with higher-difficulty challenges unattempted due to lacking older students. Dimensionality analysis confirmed a unidimensional construct—English vocabulary ability. Addressing the research question, the Rasch model helped to partially create an effective measurement scale for game challenges, evident in the robust item hierarchy differentiating challenge difficulties. However, low person separation limits its effectiveness for ability estimates. To improve the scale, recruiting more older students (12 years and above) to test harder challenges is key to increase response data and align difficulties with a broader ability range.

Keywords: Vocabulary, game design, Rasch model, English second-language learning, adaptive practice

Integrating generative artificial intelligence in higher education: A pedagogical and assessment framework review

by Jade Tan | Singapore Institute of Technology

Che Yee Lye | Singapore University of Social Sciences

Abstract ID: PROMS2025-SG002

Presenter Name(s): Jade Tan

Abstract:

With growing concerns about generative artificial intelligence (gAI) enabling plagiarism and threatening academic integrity, establishing a clear framework and guidelines for its pedagogical and assessment integration has become more critical. The past two to three years have witnessed an exponential growth in publications exploring pedagogical and assessment frameworks for integrating gAI into teaching and learning, and assessment. However, effective integration requires comprehensive theoretical considerations and thoughtful alignments between curriculum and assessment, and AI applications. This paper critically examines existing frameworks and identifies key factors affecting gAI integration in higher education. Drawing from various academic databases between 2023 and 2025, and following the EPPI systematic review approach, all relevant articles were analysed based on two levels: students and teachers. Findings revealed the need for teachers to execute learnt technical and pedagogical expertise and engage gAI as an integrative, collaborative and transformative tool in pedagogy and assessment. Teachers' openness to incorporating gAI in curriculum and their readiness are also essential in gAI integration. It was also found that students must be equipped with skills regarding AI and information literacy. Students' attitudes and perceptions towards gAI as a helpful tool and to what extent the gAI tools enhance students' interest and active participation in learning are also key factors for implementing gAI in teaching and learning. Considering the disruptive nature of gAI within the sphere of education, teachers must remain at the forefront when it comes to imparting their students with new competencies and literacies that align with the reality of modern educational and professional landscapes.

Keywords: Generative artificial intelligence, pedagogy, assessment, higher education, review

Towards a framework of multilevel analysis of student- and teacher-level factors influencing dyslexic students' reading performances

by Sharyfah Fitriya | Singapore University of Social Sciences

Che Yee Lye | Singapore University of Social Sciences

Abstract ID: PROMS2025-SG003

Presenter Name(s): Sharyfah Fitriya

Abstract:

Despite the established importance of motivation and self-efficacy on reading performance for students with dyslexia, limited research has been conducted to investigate how individual student and teacher factors affecting students' reading performance. This study adopts the Bandura's Social Cognitive Theory and proposes a multilevel analytical framework to examine the impact of student- and teacher-level factors on the English reading performance of dyslexic students studying in Singapore secondary schools. Given the nested structure of educational data where students are grouped within classrooms and teachers, the Hierarchical Linear Modelling (HLM) is employed to analyse both the student-level (Level 1) and teacher-level (Level 2) variables. HLM is selected for its strength in accounting for intra-class correlations and partitioning variance across levels. At Level 1, student predictors include motivation, self-efficacy, mindfulness practices, educational technology use, family support, and engagement. Level 2 variables comprise teacher demographics, instructional strategies, and school support systems. The Motivated Strategies for Learning Questionnaire (MLSQ), General Self-Efficacy Scale (GSE), Child and Adolescent Mindfulness Measure (CAMM) and Family Involvement Questionnaire (FIQ) will be used to collect data from students and teachers. To validate these questionnaires, Rasch analysis and Confirmatory Factor Analysis (CFA) will be conducted. The research will be conducted in two phases: a pilot study to validate the instruments within the context of this study, followed by an actual study using the validated questionnaires and Acadience standardised reading assessments. This methodological approach aims not only to produce statistically rigorous findings but also to contribute to evidence-based practices in inclusive education, especially for supporting secondary students with diverse learning needs such as dyslexia. The results will offer actionable insights for educators and policymakers seeking to optimise instructional strategies and support systems in complex, real-world classroom settings.

Keywords: Hierarchical Linear Modelling, Rasch analysis, dyslexia, reading , Bandura's Social Cognitive Theory

Developing and validating a generative AI literacy scale in postgraduates' academic writing

by Yu Liu | Nanyang Technological University

Shaoyan Zou | University of Health and Rehabilitation Sciences

Abstract ID: PROMS2025-SG004

Presenter Name(s): Yu Liu

Abstract:

The recent integration of generative AI tools into academic writing process has created an urgent need to understand and assess users' competencies in this domain. While AI literacy has emerged as a distinct research area, few studies have delved into generative AI literacy (GAIL) in the context of academic writing. Although several scales on AI literacy have been developed (e.g., Ng et al., 2024; Wang & Wang, 2025), they rarely take the nuanced demands of academic writing into consideration. Therefore, there remains a notable research gap in developing and validating a GAIL scale in the context of academic writing. This gap is particularly critical for postgraduate students who are often involved in academic writing. To address this, our study developed and validated a GAIL scale for postgraduate students' academic writing. An initial questionnaire of 35 items across five dimensions—benefits, limitations, prompts, evaluation, and ethics—was administered to 520 Chinese postgraduate students. The collected data were then submitted to a confirmatory factor analysis (CFA) using AMOS 29.0.0. To meet the composite reliability (CR) and average variance extracted (AVE) standards, we refined the scale to 24 items. The final scale demonstrated strong reliability (Cronbach's α = 0.970) and robust model fit (CMIN/df = 2.685; CFI = 0.946; RMSEA = 0.057). Path analysis revealed positive interrelationships among the five dimensions, although ethics showed comparatively weaker path coefficients, likely reflecting students' consistently high ethical awareness. Further insights were obtained by analyzing responses to an open-ended question using grounded theory. Through a three-stage coding process, 15 subcategories were distilled into five overarching themes: content quality, ethical norms, user capabilities, contextual policies, and practical utility. These findings suggest that students perceive generative AI use as a complex, interconnected phenomenon shaped by capability, policy, ethics, practicality, and content considerations. The qualitative findings provide additional explanations for the CFA results, particularly the positive relationships observed among the dimensions. The study is significant in that it conceptualized and empirically validated a GAIL scale, thus offering a valuable tool for future research and educational practice in AI-integrated academic contexts.

Keywords: Generative AI literacy scale, academic writing, confirmatory factor analysis, grounded theory

Standards-aligned authentic assessment in pharmacy technician education

by Yin Ni, Annie Ng | Nanyang Polytechnic

Cheng Keat Tan | Nanyang Polytechnic

Abstract ID: PROMS2025-SG005

Presenter Name(s): Yin Ni, Annie Ng

Abstract:

This study presents a conceptual framework for authentic assessment in pharmacy technician education, with specific application in the Medication Therapy module. The framework aligns with Singapore's Pharmacy Technicians Entry-to-Practice Competency Standards (MOH, 2022) to bridge classroom learning with professional practice requirements. Grounded in authentic assessment theories (Gulikers et al., 2004; Bloxham et al., 2017), the framework integrates five key principles: contextualized task design, cognitive complexity, iterative feedback, criterion-referenced evaluation, and learning-assessment integration. These principles guide the development of assessments that mirror actual pharmacy practice scenarios. The framework was implemented within the Medication Therapy module through a systematic approach. Students engaged in simulated practice environments encompassing prescription processing, medication preparation, and patient counseling scenarios. Assessment instruments were carefully developed through a multi-stage process involving the mapping of module learning outcomes to competency standards, design of scenario-based practical assessments, creation of analytic rubrics for performance evaluation, and implementation of formative feedback mechanisms. Application of the framework in the Medication Therapy module demonstrated several significant outcomes. It enabled structured assessment of core competencies including drug information communication and medication safety practices. The implementation achieved clear alignment between classroom tasks and professional standards, while standardized rubrics ensured objective evaluation of student performance. Furthermore, the sequenced assessment design facilitated progressive skill development, effectively preparing students for real-world pharmacy practice. This framework provides a systematic approach to authentic assessment in medication therapy training, demonstrating practical applicability in pharmacy technician education. While showing promise in standardizing competency evaluation, further research is needed to examine its transferability across different educational contexts. The study contributes to vocational education by offering a replicable model for aligning training with professional standards.

Keywords: Authentic assessment, competency-based education, pharmacy technician training, medication therapy, performance evaluation

Evaluating performance of large language models on university-level mathematics and psychology multiple-choice questions for adaptive learning system

by Hariz Zhen Wei Liew | Singapore University of Social Sciences

Che Yee Lye | Singapore University of Social Sciences

Abstract ID: PROMS2025-SG006

Presenter Name: Hariz Zhen Wei Liew

Abstract:

This study evaluates the performance of Large Language Models (LLMs) on multiple-choice questions (MCQs) in university-level mathematics and psychology, assessing their suitability for adaptive learning systems. A total of 1,111 MCQs from two mathematics and three psychology courses were analyzed, categorized by learning progression levels reflecting question complexity. Four LLMs—GPT-4o, GPT-4o-mini, GPT-o1-mini, and DeepSeek-V3—were tested for accuracy, cost-efficiency, and latency. GPT-o1-mini consistently achieved the highest accuracy in mathematics and demonstrated strong performance in psychology courses emphasizing statistical methods and reasoning tasks. However, its high computational cost and response latency present scalability challenges. GPT-4o exhibited superior performance in psychology courses, aligning with its strengths in language processing, but showed limitations in mathematical reasoning. DeepSeek-V3 displayed balanced capabilities across both subject areas, though with increased response latency. Conversely, GPT-4o-mini was cost-efficient but less accurate, particularly in mathematically complex questions. The multi-criteria evaluation revealed optimal model selection based on priority scenarios: GPT-o1-mini is recommended for mathematics where accuracy is crucial, despite higher costs. GPT-4o offers optimal performance for psychology, especially under latency-sensitive conditions, and DeepSeek-V3 provides a balanced trade-off between cost and accuracy for both disciplines. Findings underscore the importance of strategic model selection aligned with educational goals, highlighting LLMs' potential and limitations in adaptive learning environments. Limitations of the study include small sample sizes at higher complexity levels, reliance on MCQs which limit deeper cognitive assessments, evaluations restricted to specific model versions, and use of default API parameters without optimization.

Keywords: Large language models, multiple-choice questions, mathematics, psychology, adaptive learning

Pursuing a cultural understanding of distributed leadership practices among middle leaders in Singapore schools

by Simon Lim | Institute for Adult Learning, Singapore University of Social Sciences

Jonathan Goh | National Institute of Education, Nanyang Technological University

Abstract ID: PROMS2025-SG007

Presenter Name(s): Simon Lim

Abstract:

Effective school improvement requires more than just the principal's leadership. While principals are crucial, a more dispersed or decentralized approach that empowers teachers to assume expanded roles in pedagogical practices and school governance for lasting success is necessary. Western-centric notions of Distributed Leadership (DL) have gained prominence in educational research (e.g., Gronn, 2000; Spillane, 2006), however they may not be readily transferable to diverse cultural settings (Dimmock & Walker, 2004). Effective implementation of DL requires a culturally sensitive approach. Simply transferring Western models to non-Western contexts (such as Singapore) overlooks crucial cultural nuances (Collard, 2007; Hairon & Goh, 2015) and may not yield the expected benefits. The purpose of this study is to investigate the relationship between DL practices and cultural work-related values of middle leaders in Singapore schools. Two frameworks were employed to examine the relationship between the two key constructs. DL was analysed using Hairon and Goh's (2015) four-dimensional framework (empowerment, collective engagement, shared decision-making, and developing leadership), while Hofstede's (2011) six-dimensional model of cultural work values – individualism/collectivism, power distance, uncertainty avoidance, assertiveness/consideration, long-term/short-term orientation, and indulgence/restraint – would provide the cultural nuances. A total of 117 middle leaders from Singapore schools participated in this study. Ten rating scales for DL practices and cultural work values were validated and calibrated using Rasch analysis (Wright, 1993). The raw responses of middle leaders were converted to linear measures, and persons and items were placed onto common scales for measurement. Once the Persons' measures were obtained for the variables, Pearson's correlation was employed to analyse the relationship between DL practices and cultural work values dimensions. Findings reveal that leadership practices are largely related to cultural work values. Power distance was significantly but negatively correlated with empowerment, collective engagement, making shared decisions, and developing leadership. Collectivism and long-term orientation were positively correlated with collective engagement, making shared decisions, and developing leadership. Analyses of the Wright maps from Rasch analysis provided nuanced insights into the cultural work values of educators which could influence their leadership practices.

Keywords: Cultural work values, middle leaders, distributed leadership, Rasch analysis, Singapore schools

Identifying time-varying measurement model parameters in intensive longitudinal data using cross-classified factor model

by Ringo, Moon-Ho Ho | Nanyang Technological University

Jie Xin Lim | Nanyang Technological University

Abstract ID: PROMS2025-SG008

Presenter Name(s): Ringo, Moon-Ho Ho

Abstract:

This research investigated the application of the cross-classified factor model to detect time-varying measurement model parameters in intensive longitudinal data under planned-missing data design. Intensive longitudinal studies involve real-time observations of daily life which typically takes about 1 to 2 weeks, with 2 to 12 measurement occasions per day. To reduce participation fatigue, planned missing data (PMD) design is often introduced by administering subsets of items from a scale of interest, either by using the same subset (constant form) or different subsets (varying form) over time. Traditionally, testing longitudinal measurement invariance involves imposing equality constraints on the longitudinal measurement model parameters. This approach works well with a relatively small number of time points. However, this approach poses an estimation challenge with intensive longitudinal data because the number of freely estimated parameters exceeds the sample size which leads to non-positive definite variance-covariance matrices (Wothke, 1993). Recently, Muthén and Asparouhov (2012) proposed using a cross-classified factor model to model time-varying parameters as random effects to reduce the number of model parameters. However, this model alone does not identify the time-varying parameters. We proposed to adopt the algorithm proposed by Asparouhov and Muthén [A&M] (2014) to identify these time-varying/time-invariant parameters under PMD design. We used Monte Carlo simulation to investigate the accuracy of our proposed method in identifying time-varying parameters. The simulation included: 2 measurement occasions (30 and 60), 3 sample sizes (50, 200, 350), 3 autoregressive AR(1) latent factor covariance structures (ρ; 0, 0.35, 0.70), and 3 missing data designs (full scale, constant form, varying form). A unidimensional factor model with five continuous indicators was used to simulate the intensive longitudinal data with time-varying parameters. Each condition was replicated 500 times. Accuracy in identifying time-varying parameters decreased when missing data were present but the impact diminished with larger sample sizes and more measurement occasions. The accuracy in identifying time-varying loadings was also affected by the covariance structure, particularly with small sample sizes and fewer measurement occasions. The results demonstrated that the cross-classified factor model, combined with A&M's (2014) algorithm can be reliably used to detect measurement invariance in the presence of planned missing data.

Keywords: Planned-missing data, measurement Invariance, psychometrics

Guessing as ability rather than item characteristic: A new framework for mixture Item Response Theory

by Metin Bulus | Adiyaman University

Abstract ID: PROMS2025-TR001

Presenter Name(s): Metin Bulus

Abstract:

The primary objective of this study is to propose and evaluate a novel extension to Item Response Theory (IRT) that re-conceptualizes guessing as a latent trait inherent to individuals, rather than as a static property of test items. Traditional IRT frameworks—especially the widely used three-parameter logistic (3PL) and four-parameter logistic (4PL) models—treat guessing as an item-level characteristic, typically captured by a lower asymptote parameter) representing the probability that a respondent would guess the correct answer. This assumption, however, overlooks individual variability in guessing behavior, such as differences in risk aversion, test-wiseness, or confidence, which are rooted in personal traits rather than item properties. In addition to relocating the guessing parameter from items to individuals, the model also seeks to leverage valuable, often underutilized information embedded within distractors—the incorrect response options in multiple-choice items. While classical IRT models focus solely on whether a response is correct or incorrect, distractors may provide diagnostic insight into respondents' partial knowledge or misconceptions. This research proposes a unified framework that integrates these two dimensions, aiming to enhance the interpretability and precision of trait estimates in educational and psychological assessments.

Keywords: Mixture item response theory, guessing, distractor analysis

A practical guide to sample size calculations in psychometric research

by Metin Bulus | Adiyaman University

Abstract ID: PROMS2025-TR002

Presenter Name(s): Metin Bulus

Abstract:

Determining an adequate sample size is critical to ensuring robust, reliable, and valid results in psychometric research. Sample size decisions affect the stability of statistical estimations, the accuracy of parameter estimates, and the overall generalizability of findings. However, researchers often overlook systematic approaches in favor of convenience or tradition, leading to either excessively large samples (resulting in unnecessary resource expenditure) or insufficiently small samples (risking unreliable findings). Miscalculations and misconceptions regarding sample size can seriously undermine psychometric quality, potentially resulting in misleading conclusions (Brown, 2015; Wolf et al., 2013). This practical guide aims to clarify the complexities surrounding sample size calculations in psychometric contexts. By offering clear, actionable recommendations and illustrative examples, this paper will assist researchers in making informed decisions tailored specifically to psychometric methodologies, including scale development, reliability assessments, and factor analyses.

Keywords: Power analysis, sample size, psychometric

Differential item functioning analysis in PISA 2022 using Rasch trees: Finland, Turkey, and Singapore

by Enes Yavuz | Gebze Technical University

Abstract ID: PROMS2025-TR003

Presenter Name(s): Enes Yavuz

Abstract:

This study explores whether Differential Item Functioning (DIF) related to cross-cultural differences exists among students from Finland, Turkey, and Singapore who took part in PISA 2022. These countries were chosen because Finland ranks highest in Europe, Singapore leads globally and in Asia, and Turkey serves as a cultural bridge between Asia and Europe. The sample includes about 24,000 students. Since not all students answered the same test booklets, only the booklets common to all three countries were analyzed. DIF happens when test questions perform differently for groups with similar ability, suggesting potential bias. To detect this, Rasch measurement theory combined with Rasch trees was used, which allows to dig deeper into fairness across culturally diverse groups. The official OECD PISA 2022 database was used, including both student questionnaire responses and cognitive test items. Before starting the DIF analysis, missing data and descriptive statistics were checked to get a clear picture of data quality. Unlike traditional methods like Mantel-Haenszel or Logistic Regression, which require defining groups in advance, Rasch tree method takes a more flexible, data-driven approach. The analyses were performed using raschtree function in R's psychotree package. By examining Rasch tree graphs, including nodes and parameter estimates, several items were identified that showed significant DIF across the three countries—indicating real cultural differences in how questions function beyond students' abilities. Also, gender-related DIF was found, with some questions favoring boys or girls, and this was consistent across all three countries. These results highlight how important it is to consider cultural and gender factors in international tests like PISA to keep assessments fair and meaningful. While Rasch models are standard in scoring PISA, Rasch trees give a more detailed lens to spot and understand DIF. One limitation is that this study only looked at three countries, so expanding this work to more nations would help make the findings even stronger. Overall, Rasch trees prove to be a powerful tool for uncovering hidden biases and ensuring valid cross-cultural comparisons.

Keywords: Differential Item Functioning (DIF), Rasch Measurement Theory, Rasch trees, PISA 2022, cross-cultural comparison

Developing a revised DIF-free-then-DIF strategy to simultaneously assess uniform and nonuniform DIF

by Wei-Chia Su | National Sun Yat-sen University

Po-Hsien Hu | National Sun Yat-sen University

Ching-Lin Shih | National Sun Yat-sen University

Abstract ID: PROMS2025-TW001

Presenter Name(s): Wei-Chia Su

Abstract:

The presence of differential item functioning (DIF) items can bias parameter estimates in item response models. A strategy known as "DIF-free-then-DIF" has been shown to yield better-controlled Type I error rates and higher power rates than traditional methods when assessing uniform DIF. This strategy first identifies a set of DIF-free anchor items to serve as the matching variable, followed by DIF assessment using the constant-item method. Given that both uniform and nonuniform DIF have been observed in prior research—and that each type may influence parameter estimates differently—a revised DIF-free-then-DIF strategy capable of assessing both types of DIF simultaneously is needed. Logistic regression is a flexible method that can simultaneously assess both uniform and nonuniform DIF. In this study, two approaches—one-step and two-step—were compared under the two-parameter logistic (2PL) model through a series of simulation studies. It was hypothesized that the one-step approach would perform as well as or better than the two-step approach. Four independent variables were manipulated: (a) DIF assessment method: one-step vs. two-step; (b) sample size combinations: R250/F250, R500/F250, R500/F500, and R1000/F500, where R and F denote the reference and focal groups, respectively; (c) proportion of DIF items in the test: 0%, 10%, 20%, 30%, and 40%; and (d) type of DIF among DIF items: all uniform, all nonuniform, or a 50/50 mix. The dependent variables included: (a) the accuracy of identifying DIF-free items, (b) Type I error rate, and (c) power rate. The one-step approach demonstrated higher accuracy and greater power than the two-step approach while maintaining well-controlled Type I error rates. Moreover, it exhibited higher computational efficiency (i.e., shorter processing time). Both uniform and nonuniform DIF are commonly found in real-world assessments, particularly in international large-scale assessments. To ensure the reliability and validity of test scores, both types of DIF should be examined. This study proposed a revised DIF-free-then-DIF strategy based on logistic regression that can simultaneously assess both types of DIF. Preliminary simulation results suggest that this approach offers a more efficient and effective solution for DIF assessment.

Keywords: Differential item functioning, logistic regression, DIF-free-then-DIF, uniform DIF, nonuniform DI

Development and validation of a framework for assessing linguistic competencies in senior-year Chinese majors

by Suet Ching Soon | National United University

Chia-Ling Hsu | Hong Kong Examinations and Assessment Authority

Abstract ID: PROMS2025-TW002

Presenter Name(s): Suet Ching Soon, Chia-Ling Hsu

Abstract:

Syntactic knowledge is a fundamental element of linguistic competence and plays a vital role for those planning to teach or work in language-focused professions. This study aims to develop a 40-item instrument for measuring syntactic knowledge in Chinese, based on six basic syntactic structures (including Subject-Predicate Structure, Modifier-Head Structure, Verb-Complement Structure etc.), and to examine its psychometric properties through Rasch analysis using a sample of 112 undergraduate students enrolled in the Department of Chinese Language and Literature at National United University in Taiwan. To enhance participants' motivation and engagement during the assessment, test items were organized with easier items placed at the beginning and progressively more difficult ones towards the end. The items were randomly selected while ensuring no option appeared consecutively more than three times. Each item was scored as 1 for a correct response and 0 otherwise. The Rasch analysis yielded the following findings: (a) all 40 items effectively measured the intended general construct of syntactic knowledge in Chinese sentential comprehension; (b) the fit indices (unweighted mean square error and weighted mean square error) indicated a good model-data fit as their values fell within the utilized criterion of 0.7–1.3 (Linacre, 2006 ; Wright & Linacare, 1994); (c) the person separation reliability (PSR) demonstrated excellent reliability, with a value exceeding .99 ; and (d) more than 95% of the item difficulty values were below -1.0 logit. In sum, the Rasch analysis validated the psychometric properties of the 40-item instrument with respect to both construct validity and reliability at the item level. The results support the use of this tool for assessing students' syntactic knowledge in Chinese. Additionally, the fact that most item difficulty values were lower than -1.0 logit suggests its potential as a diagnostic tool for classroom assessments, particularly for identifying individuals with low to moderate ability levels and informing instructional adjustments or targeted interventions.

Keywords: Rasch Analysis, instrument validation, Chinese assessment, linguistic competency, syntactic knowledge

Development of a multidimensional mathematical competence adaptive test: Item bank construction, simulation, and empirical analysis

by Yu-Chun Lien | National Taiwan Normal University

Yao-Ting Sung | National Taiwan Normal University

Wei-Hung Yang | National Taiwan Normal University

Abstract ID: PROMS2025-TW003

Presenter Name(s): Yu-Chun Lien

Abstract:

This study employed Item Response Theory (IRT) methodologies to develop and validate the Mathematics Competence Assessment and Diagnosis (MCAD) system—a multidimensional computerized adaptive test (MCAT) designed to assess students' mathematical abilities across diverse domains. The goal was to create an assessment that is both precise and efficient while supporting adaptive instruction and real-world educational applications. However, most existing assessments rely on unidimensional models or fixed-form testing, which limit measurement precision and increase the testing burden. To address these limitations, this study was conducted in three phases to ensure the psychometric quality of the system and its alignment with instructional practices. First, an item bank was developed through a five-step process: item construction, booklet design, participant sampling, test administration, and statistical analysis. A total of 316 items were retained, covering four mathematical dimensions—Quantity, Space and Shape, Change and Relationships, and Uncertainty and Data—based on Taiwan's national curriculum and the OECD's PISA framework. Second, a simulation study compared three item selection strategies: MCAT, Unidimensional CAT (UCAT), and Random Administration (RA). The MCAD system using MCAT consistently outperformed both UCAT and RA in measurement precision (demonstrated by lower RMSE) and test efficiency, achieving high reliability with significantly fewer items. Third, an empirical study involving 105 tenth-grade students validated the system's effectiveness. MCAD scores showed a strong correlation (r = .70) with students' mathematics performance on the Comprehensive Assessment Program (CAP), confirming criterion-related validity. ANOVA results further demonstrated that MCAD could significantly differentiate among high-, medium-, and low-performing students. The MCAD system provides real-time, personalized assessment feedback while reducing testing burden, making it suitable for both classroom-based assessments and large-scale standardized testing. It supports school teachers in obtaining precise and efficient adaptive measurement results. This study illustrates the potential of integrating multidimensional IRT and adaptive testing to enhance the measurement precision and efficiency of mathematics assessments.

Keywords: Item response theory, computerized adaptive testing, mathematical competence test, multidimensional analysis, testing efficiency

Predicting IRT-based word difficulty using deep neural networks: A semantic feature-based approach

by Wei-Hung Yang | National Taiwan Normal University

Yao-Ting Sung | National Taiwan Normal University

Yu-Chun Lien | National Taiwan Normal University

Chia-Hsin Chen | National Taiwan Normal University

Abstract ID: PROMS2025-TW004

Presenter Name(s): Wei-Hung Yang

Abstract:

This study explores the feasibility of applying machine learning techniques to predict word difficulty levels and proposes an automated framework for test development based on psychometric modeling. The participants included 210 students in Taiwan, ranging from third to ninth grade. Each participant completed a fill-in-the-blank vocabulary test involving 1,830 English words, for which they provided the corresponding Chinese meanings. Their responses were used to estimate word difficulty parameters using the one-parameter item response theory (1PL IRT) model. Subsequently, 33 semantic features were extracted for each word, including word frequency, semantic abstractness, and semantic distance. These features were then used to train a deep neural network (DNN) to learn the mapping between semantic characteristics and IRT-based difficulty estimates. The model achieved a prediction accuracy of 89.3%, demonstrating high performance in estimating word difficulty. This study provides empirical evidence of the relationship between semantic features and word difficulty. It also shows the potential of machine learning in language test development, offering a pathway for automating item construction, reducing the resources required for traditional test design, and improving the efficiency and precision of second language assessment.

Keywords: 1PL IRT, semantic features, machine learning, deep neural network, word difficulty prediction

Recovering person, item and dispersion parameters in the extended continuous Rating scale model

by Yeh-Tai Chou | National Taiwan Normal University

Yao-Ting Sung | National Taiwan Normal University

Pin-Hsun Song | National Taiwan Normal University

Abstract ID: PROMS2025-TW005

Presenter Name(s): Yeh-Tai Chou

Abstract:

The extended continuous rating scale model (Verhelst, 2019), a member of the Rasch model family, was developed to accommodate item responses collected using continuous rating formats, such as the visual analogue scale. However, its application in research and practice remains limited. To promote broader adoption of the model, a reliable and efficient parameter estimation algorithm is essential. The primary aim of this study was to develop and evaluate an estimation algorithm for recovering the model parameters—including person abilities, item difficulties, and item-level dispersions—of the extended continuous rating scale model. A series of Markov chain Monte Carlo (MCMC) simulations were conducted to assess the performance of the proposed algorithm in terms of parameter recovery. Simulation conditions varied according to sample size (200, 500, and 1,000 examinees) and test length (20, 40, and 60 items). Person ability parameters were sampled from a standard normal distribution, N(0, 1). Item difficulty parameters were uniformly distributed from –2.00 to 2.00, and item-level dispersion parameters were sampled from a uniform distribution ranging from 0.30 to 2.00. Item responses were simulated as continuous scores within the range [0.0, 5.0], approximating the structure of a traditional 5-point Likert scale. Each simulated dataset was calibrated using the proposed estimation algorithm. The accuracy of parameter recovery was evaluated by computing the mean absolute deviation (MAD) and root mean square error (RMSE) across 100 replications per condition. The results indicated that MAD values for person ability estimates were below 0.29, and RMSE values were below 0.36 across all conditions, even with as few as 200 examinees and 20 items. Estimation accuracy was even higher for item difficulty and dispersion parameters, demonstrating the robustness of the algorithm across varying conditions. In summary, the proposed estimation algorithm can accurately estimate parameters of the extended continuous rating scale model. The resulting person and item estimates can be treated as interval-level scores, thus fulfilling the assumptions required for applying parametric statistical analyses in subsequent research.

Keywords: Continuous data, visual analogue scale, Rasch model, estimation algorithm, parameter recovery

A mixed Rasch modelling approach to investigating teacher resilience in Malaysia

by Zhi Jie Lee | The Ohio State University

Sharifah Hanizah Syed Jaafar | Ministry of Education Malaysia

Esther Tan | Institute of Teacher Education Ilmu Khas Campus

Mei Ai Foo | Wangsa Maju Chinese National-Type Primary School

Abstract ID: PROMS2025-US001

Presenter Name(s): Zhi Jie Lee, Sharifah Hanizah Syed Jaafar, Esther Tan

Abstract:

This study examined the latent structure of teacher resilience among Malaysian educators using the Mixed Rasch Model (MRM; Rost, 1990). The objective was to identify distinct subgroups and evaluate the psychometric performance of resilience-related items within each group. The study was grounded in resilience theory and person-centred measurement approaches. It hypothesised the emergence of multiple latent classes within the teaching workforce, each reflecting a unique resilience profile. The MRM was chosen to account for heterogeneity in item response patterns and to provide class-specific psychometric insights. Participants included 2,324 Malaysian teachers who completed a ten-item instrument designed to measure key indicators of teacher resilience. Items were rated on a 4-point Likert scale ranging from “Strongly disagree” (1) to “Strongly agree” (4). MRM analysis was conducted to determine the optimal number of latent classes and to evaluate item properties such as item difficulty, item fit, item polarity, and item reliability within each class. Results supported a two-class solution as the best fit to the data, indicated by lower Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values relative to other models. The two classes demonstrated distinct item difficulty hierarchies. The result suggested the presence of both common and divergent resilience patterns among Malaysian educators. Item-level analyses revealed acceptable psychometric properties across classes. These findings underscore the existence of meaningful subgroups within the teaching population and highlight the importance of differentiated, data-informed approaches to supporting teacher resilience. The study contributes to both measurement and practice by validating a culturally relevant resilience instrument and identifying targeted support needs within Malaysia’s educational context.

Keywords: Teacher resilience, Malaysian teachers, mixed Rasch model, latent class analysis

School bullying victimisation in Malaysia: A mixed Rasch model approach for school counselling

by Zhi Jie Lee | The Ohio State University

Mei Ai Foo | Wangsa Maju Chinese National-Type Primary School

Esther Tan | Institute of Teacher Education Ilmu Khas Campus

Abstract ID: PROMS2025-US002

Presenter Name(s): Esther Tan

Abstract:

This study utilised the Mixed Rasch Model (MRM; Rost, 1990) to identify latent classes of school bullying victimisation experiences among secondary school students in Malaysia and to evaluate the psychometric properties of the items within each latent class. Based on previous research across different countries and cultures, it was hypothesised that multiple latent classes exist, representing distinct patterns of victimisation. Using the 2022 public dataset from the Programme for International Student Assessment (PISA), the study analysed responses from 7,069 Malaysian students, aged 15 years, across 199 secondary schools. Six indicators of peer victimisation were examined: (i) deliberate exclusion, (ii) mockery, (iii) threats, (iv) property taken or destroyed, (v) physical attacks, and (vi) rumours spread. Each indicator was assessed on a 4-point scale, where a score of “1” represented “Never or almost never” and a score of “4” indicated “Once a week or more,” capturing the frequency of bullying experiences. Results revealed that a four-class model provided the best fit, as determined by lower Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values compared to alternative models. The identified latent classes were: (i) Excluded and Materially Victimised, (ii) Excluded and Physically Victimised, (iii) Threatened Target of Rumours, and (iv) Threatened Physical and Property Bullying. These classes demonstrated varying item difficulty ordering and indicated both shared and distinct victimisation experiences. By examining these latent classes and the psychometric properties (i.e., item fit, item polarity, and item reliability) of the items within each class, the study proposes comprehensive school counselling strategies tailored to students’ diverse needs. The findings highlight the necessity for multi-tiered interventions, including schoolwide prevention efforts, classroom lessons, small group counselling, and targeted individual counselling initiatives. These interventions should focus on promoting protective factors such as empathy, compassion, and prosocial behaviours to mitigate the negative impacts of bullying victimisation in school settings.

Keywords: School bullying victimisation, Malaysian secondary schools, school counselling, mixed Rasch model, latent class analysis

Development of an eleven-item scale for measuring food insecurity

by Jing Li | University of Georgia

George Engelhard Jr. | University of Georgia

Abstract ID: PROMS2025-US003

Presenter Name: Jing Li

Abstract:

The purpose of this study is to develop an eleven-item scale for measuring food insecurity based on a subset of items used in the Household Food Security Survey Module (United States Department of Agriculture; USDA). A polytomous Rasch model is used to calibrate the scale. The data are based on families with children who participated in Current Population Survey Food Security Supplement (CPS-FSS) in 2019 in the United States (N=1,248). This study differs from previous research in several ways. First of all, a continuous scale based on Rasch measurement theory is developed. A continuous scale provides increased sensitivity to changes in the severity of food insecurity as compared to simply reporting food insecurity in categories. Another feature of the new scale is that the content alignment between child and adult items on the scale. Also, our approach uses the ratings directly obtained from respondents based on a rating scale structure rather than dichotomizing items. Overall, the model fit the data very well with 64.2 percent of the variance explained. The scale also provides the opportunity to determine cut scores so that households can be assigned to USDA food insecurity categories. A scoring table is provided to convert observed scores to a metric scale based on the Rasch model. This study has implications for research, theory and practice related to the measurement of food insecurity as well as secondary analyses of food insecurity.

Keywords: Rasch models, food insecurity, polytomous data, model-data fit

PROMS2025 Sponsors

STALL Sponsor

CommonTown brings over a decade of Rasch Model expertise through R&D and application in adaptive learning platforms and education management systems. Our successes include language placement testing, Chinese and English adaptive reading systems, item bank calibration, and optimized deliberate practice algorithms. We are also pioneering AI-driven tutoring and automated learning game creation embedded with Rasch Model.

STUDENT Sponsor

THANK YOU

proms@suss.edu.sg

proms2025.com

463 Clementi Road, Singapore 599494