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

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.

[PROMS2025-CN004] Exploring the moments of insight in human-AI co-creative process

Author(s): Sujie Yang, Manli Zhang, Jue Wang

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.

[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

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.

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

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.

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

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.

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

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.

[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

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.

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

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.

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

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.

[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

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.

[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

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.