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

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.

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

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

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.

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

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.

[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

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.

[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

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.

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

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.

[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

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.

[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

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.