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

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.

[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

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.

Theme: Measurement Theory & Practice

Venue: SR.C.3.15

Chair: Dr Evelyn Gay

[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

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.

[PROMS2025-AU002] Optimizing test length in assessments through adaptive simulation

Author(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.

[PROMS2025-HK004] Variable-length multistage adaptive testing design

Author(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.

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

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.

[PROMS2025-PH003] Development and Rasch analysis of the critical thinking test in chemistry

Author(s): Jonathan Barcelo, 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.

[PROMS2025-PH002] Development and validation of the Visual Representations Test using Rasch measurement framework

Author(s): Jonathan Barcelo, 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.