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.

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.

[PROMS2025-PH004] Application of Rasch analysis in the evaluation of biochemistry examination for health science students

Author(s): Jonathan Barcelo, 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.

[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

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.

[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

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.

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

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.

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

[PROMS2025-TR002] A practical guide to sample size calculations in psychometric research
Author(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.

[PROMS2025-IN007] How to measure financial intelligence: Rasch model analysis
Author(s): Muhammad Rayhan I'tisham, Tutut Chusniyah, Ninik Setiyowati, Kukuh Setyo Pambudi, Hariss Shah Abd Hamid, Hadi Sumarsono

Abstract:
Financial intelligence has so far been defined from an economic perspective, which causes overlap with financial literacy. Financial intelligence should include the definition and terminology of "intelligence" and "finance" from a psychological perspective, not just financial knowledge. The research method used is quantitative psychometric analysis to develop a measuring tool with rasch model analysis. In the quantitative study of the development of measuring instruments, the first pilot testing (N=355) was carried out for the first calibration item, especially on the bias of the domain, and the second pilot testing (N=208) for the second calibration item, especially on the refinement of validity, reliability, item differential power, and other psychometric parameters. The results showed good psychometrics with 62 items. Summary statistics show good psychometric properties. Matching items and people and clothing show good results at the threshold. MNSQ and ZSTD also show optimal values. The separation of items and people is also good with good item reliability. This scale is reliable and valid. Further research can use financial intelligence measures to explore the model and its relationship to other psychological variables.

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

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.

[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

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.

[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

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.

[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

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.