JP3I (Jurnal Pengukuran Psikologi dan Pendidikan Indonesia)
DOI: http://dx.doi.org/10.15408/jp3i.v14i2.42886
http://journal.uinjkt.ac.id/index.php/jp3i

Comparing IRT Models: Summated Scaling Effects on Critical Thinking in Vocational Students

Andi Abdurrahman Manggaberani[1], Samsul Hadi[2], Nur Hidayanto Pancoro Setyo Putro[3], Abrar Syahrul Fajri[4], Heri Retnawati[5]

Department of Educational Research and Evaluation, Graduate School, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia[1,2,3,4,5]

andiabdurrahman.2023@student.uny.ac.id

Abstract

This study investigates the comparative efficacy of Summated Rating Scales (SRS) and traditional ordinal scales (raw Likert-type responses) in measuring critical thinking skills among vocational students, employing Item Response Theory (IRT) to evaluate their psychometric properties. Addressing the limitations of ordinal scales notably inconsistent intervals between response categories the research adopts a descriptive quantitative methodology involving 269 students from state vocational high schools in Yogyakarta, Indonesia. Data were collected using a five-point Likert scale instrument, validated for content (Aiken's V = 0.94), and analyzed through two IRT frameworks: Polytomous IRT for unscaled ordinal data and Continuous Response Model (CRM) IRT for SRS-transformed interval data. Key findings reveal that SRS enhances measurement precision by normalizing response distributions into proportional intervals (e.g., recalibrated scores: 0.00, 0.73, 1.46, 2.07, 2.84), thereby resolving issues of unequal category spacing inherent to ordinal scales. Polytomous IRT demonstrated robust item fit (e.g., Partial Credit Model fit for 5/6 items) and strong difficulty parameter invariance (r = 0.84), yet exhibited instability in ability estimates (r = 0.37) due to extreme response patterns. Conversely, CRM IRT applied to scaled data produced stable ability estimates (r = 0.46) and eliminated infinite values in Maximum Likelihood Estimation, underscoring its superiority in handling continuous metrics. However, ordinal scales retained higher consistency in difficulty calibration across subgroups. The study concludes that integrating SRS with CRM IRT offers a refined approach for critical thinking assessments, balancing precision and fairness, while ordinal scales remain pragmatic for contexts prioritizing simplicity. These insights advocate for the adoption of advanced scaling techniques in vocational education to improve the validity of competency evaluations, with recommendations for future research to explore hybrid models and longitudinal applications.

Keywords: critical thinking, item response theory, IRT polytomous, continuous response model, summated scaling.

Abstrak

Penelitian ini mengkaji perbandingan efektivitas Skala Penilaian Terjumlah (SRS) dan skala ordinal tradisional (respons skala Likert mentah tanpa transformasi) dalam mengukur keterampilan berpikir kritis siswa kejuruan, menggunakan Teori Respons Butir (IRT) untuk mengevaluasi properti psikometri nya. Menanggapi keterbatasan skala ordinal khususnya ketidakkonsistenan interval antar kategori respons penelitian ini menggunakan metodologi kuantitatif deskriptif dengan melibatkan 269 siswa dari SMK negeri di Yogyakarta, Indonesia. Data dikumpulkan melalui instrumen skala Likert lima poin yang divalidasi konten (koefisien V Aiken = 0,94) dan dianalisis dengan dua model IRT: IRT politomus untuk data ordinal tidak terukur dan Model Respon Kontinu (CRM) IRT untuk *data interval hasil transformasi SRS. Temuan utama menunjukkan bahwa SRS meningkatkan presisi pengukuran dengan normalisasi distribusi respon menjadi interval proporsional (misal, skor terkalibrasi: 0,00; 0,73; 1,46; 2,07; 2,84), mengatasi masalah ketidaksamaan jarak kategori pada skala ordinal. IRT politomus menunjukkan kecocokan item yang kuat (misal, Model Kredit Parsial cocok untuk 5/6 item) dan invariansi parameter kesulitan yang tinggi (r = 0,84), namun menghasilkan estimasi kemampuan yang tidak stabil (r = 0,37) akibat pola respon ekstrem. Sebaliknya, CRM IRT pada data terukur menghasilkan estimasi kemampuan stabil (r = 0,46) dan menghilangkan nilai tak hingga dalam Estimasi Kemungkinan Maksimum, menegaskan keunggulannya dalam menangani data kontinu. Meski demikian, skala ordinal mempertahankan konsistensi kalibrasi kesulitan antar subkelompok. Studi ini menyimpulkan bahwa integrasi SRS dengan CRM IRT menawarkan pendekatan lebih akurat untuk asesmen berpikir kritis, menggabungkan presisi dan keadilan, sementara skala ordinal tetap praktis dalam konteks yang mengutamakan kesederhanaan. Temuan ini mendorong adopsi teknik penskalaan lanjutan dalam pendidikan kejuruan untuk meningkatkan validitas evaluasi kompetensi, dengan rekomendasi penelitian lanjutan untuk mengeksplorasi model hibrid dan aplikasi longitudinal.*

***Kata kunci**: berpikir kritis, model respon kontinu, penskalaan summated rating, teori respon butir, teori respon butir politomus.*

Introduction

The assessment of various human traits, conditions, and disorders often begins with a series of items in a questionnaire. The response format for these items is typically binary (e.g., yes/no) or ordinal (e.g., "strongly agree," "agree," or "disagree") (Casper et al., 2020). This approach facilitates the measurement of diverse variables that reflect an order or hierarchy in individuals' response levels. Ordinal scales are frequently employed across disciplines such as educational research, psychology, and social sciences due to their ability to classify responses in an order that indicates the intensity or level of agreement. However, despite their utility in providing an overview of rankings, ordinal scales have a fundamental limitation: The inconsistency in the intervals between categories. This inconsistency poses challenges for conducting more in-depth and precise statistical analyses.

The primary limitation of the ordinal scale lies in the fact that, although it can represent an order, it often fails to measure consistent distances between response categories. This can potentially reduce the accuracy and validity of more complex statistical analyses, such as those assessing students' critical thinking skills. Critical thinking skills are essential competencies that must be developed in vocational education, focusing on enhancing analytical abilities and problem-solving across various fields. Therefore, it is crucial to have accurate and consistent measurement instruments for evaluating these skills. The use of an ordinal scale in measuring critical thinking abilities may yield results that are not entirely valid, especially if the collected data fails to provide sufficient information about the differences between the various levels of skill proficiency.

To address this issue, a more effective approach is required to transform ordinal data into interval data that is more reliable for complex statistical analyses. One effective method to tackle this challenge is the Summated Rating Scale (SRS). By employing a transformation process, SRS converts ordinal data into interval data by mapping each variable onto a normal distribution, thereby ensuring consistency across categories and producing more precise scores (Kusmaryono et al., 2022). This process involves calculating frequencies, cumulative proportions, densities, and adjusted z-scores, resulting in scores that more accurately represent the variables or constructs being measured (Zou et al., 2023). The transformation procedure begins with calculating the frequency (f) of responses/categories, followed by converting them into proportions (p) and determining cumulative proportions (cum). Next, the density is calculated (Density = pk - 0.5 * p), after which z-scores are derived using the normal distribution. Finally, adjustments are made with z* to assign the scores (Febriana & Setiawati, 2024).

The application of the SRS method in measuring critical thinking skills has significant implications for vocational education (Ebil et al., 2020; Mustika et al., 2019). However, while traditional ordinal scales are often utilized due to their simplicity, their primary limitation lies in the inability to provide consistent intervals between categories. In contrast, the SRS method offers greater flexibility by improving the consistency of intervals, resulting in more stable and valid measurements (Alamrani et al., 2023; Tsikritsis et al., 2022). Traditional ordinal approach refers to the use of raw Likert-type responses analyzed at their original ordinal level, without any transformation into interval data. In this approach, each response category is assumed to be ordered but not equally spaced, and analyses are performed using statistical or psychometric methods that preserve the ordinal nature of the data. Despite the recognized potential of both scales, comparative studies focusing on their application in measuring critical thinking skills remain scarce (Shaw et al., 2020; Tobón & Luna-nemecio, 2021).

A more contemporary approach to analyzing measurement instruments involves the use of Item Response Theory (IRT) (Van Hauwaert et al., 2020; Zarate et al., 2023), which enables a more nuanced evaluation of each item within the instrument. IRT is a more advanced method compared to classical techniques, as it allows for detailed assessments of item fit, parameter estimation, and measurement efficiency (Alordiah & Oji, 2024; Selçuk & Demir, 2024). In the context of this study, IRT will be employed to compare the data generated through the scaling and without scaling process in measuring students' critical thinking skills.

The application of IRT in this study plays a critical role in evaluating the effectiveness of ordinal data (not scales) and the data that has been scaled using Summated Rating (SR) in measuring students' critical thinking skills. However, data or items that have been transformed using the SR method in the context of IRT can be considered as continuous response data. Most of the analyses conducted have largely been framed within the Classical Test Theory (CTT) framework, which distinguishes between true scores, essentially expected responses, and error scores (Lord & Novick, 2008). This study aims to explore how original ordinal data and SR-scaled data perform within the IRT framework. The rationale is to determine whether SR transformation aligns with IRT assumptions and enhances measurement precision. IRT models the relationship between item responses and latent traits non-linearly, offering more nuanced insights than CTT. However, using SR scaled data in IRT raises questions about whether it preserves the data's ordinal nature or introduces distortions.

IRT facilitates a more comprehensive evaluation of model fit, item fit, and parameter invariance analysis, all of which are essential for assessing the extent to which a measurement model can be consistently applied across diverse groups of students with varying characteristics and backgrounds (Akour et al., 2022; Dai et al., 2021; Robitzsch & Lüdtke, 2020). A significant advantage of IRT lies in its ability to evaluate the quality of individual items within an instrument, enabling the identification of items that may function inconsistently or unfairly across different levels of student ability (Jebb et al., 2021; Vollmer & Alkire, 2022). This process is crucial for ensuring that each item used in the measurement instrument accurately reflects the intended construct and provides valid information for assessing critical thinking skills.

Additionally, IRT facilitates the comparison of parameter estimations across different scales. These parameter estimations include the discrimination parameter and the difficulty parameter (Robitzsch, 2021; Şad, 2020). The comparison between SRS and ordinal scales in terms of these parameter estimations provides a clearer understanding of how effectively each scale reflects differences in students' critical thinking abilities. Consequently, IRT not only assesses model fit but also enables a more detailed comparison of how efficiently and accurately each scale conveys information about critical thinking skills.

Previous studies, such as those conducted by Mohamadi (2018), Kinel et al. (2021), Kadigi et al. (2023), Lindner and Lindner (2024) and Astuti et al. (2024), have extensively discussed the comparison between original (ordinal) scores and standardized (summated rating) scores. However, to date, there is a limited body of research that compares these two types of scores using the Item Response Theory (IRT) approach. The analysis results derived from the application of item response theory (IRT) are crucial for further discussions in this study, particularly in identifying the strengths and weaknesses of the data generated through the scaling and without scaling process by integrating IRT.

This study aims to examine and compare the impact of using interval data obtained through scaling with ordinal data without scaling in measuring the critical reasoning ability of vocational students. The focus is on students from various vocational study programs who have participated in Merdeka Curriculum-based assessments. The sample consisted of 269 students selected through purposive sampling from four state vocational high schools (SMKN) in Yogyakarta City during the 2024/2025 academic year. These schools were chosen based on program diversity, accreditation status, and comprehensive implementation of the Merdeka Curriculum, while ensuring representation from a range of expertise areas to capture variation in educational backgrounds and competency levels. This approach allows the research to obtain a comprehensive picture of the critical thinking skills of SMK students from diverse educational backgrounds and competency levels combined with IRT comparative analysis. The analysis was conducted through two approaches: first, using the polytomous IRT model on raw data from instruments without scaling; and second, using the Continuous Response Model (CRM) model on data that had been processed with the Summated Rating (SR). The two procedures are then compared to ensure a valid and meaningful comparison between the two approaches.

Methods

The research utilizes a descriptive quantitative approach to analyze and outline the key characteristics of the compiled dataset (Fialho & Zyngier, 2023; Kadim & Sunardi, 2021; Sidel et al., 2018). This study involved 269 students from state vocational high schools (SMK) in Yogyakarta City during the 2024/2025 academic year. The sample was selected using purposive sampling, a non-probability technique that allows researchers to deliberately select participants with specific characteristics relevant to the research objectives. The selection was based on data from Yogyakarta City's new student admission records and the identification of schools that had comprehensively implemented the Merdeka Curriculum. The four SMKNs were chosen to reflect diversity in program offerings, accreditation status, and student population. Within each school, students from different study programs were included to ensure variation, not to reflect proportionality statistically. This approach allowed the study to capture a comprehensive perspective on critical thinking skills, so the findings are relevant to different educational backgrounds and competency levels. The distribution of the sampled students is presented in Table 1.

Table 1. Distribution of Sampled Students from State Vocational High Schools in Yogyakarta City

School Name	Number of Students (F)	Percentage (%)
SMK A	22	8.21%
SMK B	45	16.79%
SMK C	37	13.81%
SMK D	164	61.19%
Total	269	100%

Sources: Personal data (2024).

This distribution was designed to support data analysis while maintaining proportional representation from each school, ensuring the study results could be interpreted comprehensively and relevantly. The components and subcomponents of these dimensions are detailed and mapped in Table 2.

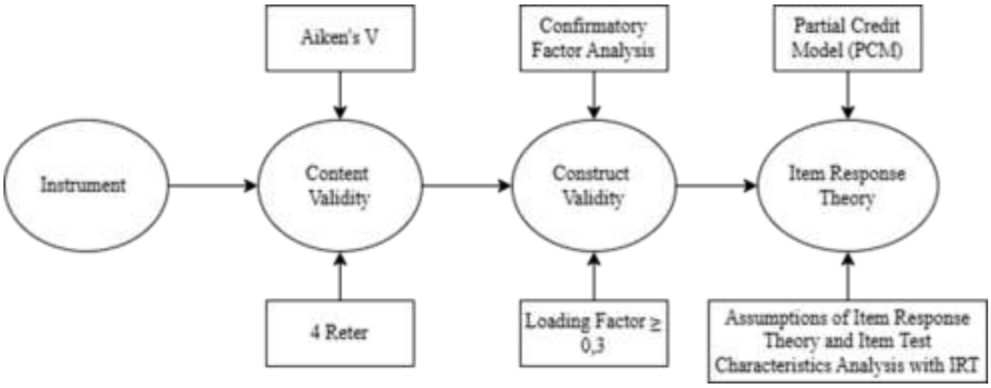
Table 2. The Distribution of Indicators and Items

Aspect/ Dimension	Indicators/Elements	Sub-indicators/Sub-elements	Number of Items	Code
Critical Thinking	Obtaining and processing information and ideas	Posing inquiries	1	Item_1
		Identifying, clarifying, and processing information and ideas	2	Item_2, Item_3
	Analyzing and evaluating reasoning and procedures	Analyzing and evaluating reasoning and its procedures	2	Item_4, Item_5
	Reflecting on thoughts and cognitive processes	Reflecting on and assessing one's own thinking	1	Item_6

Sources: Personal data (2024).

The questionnaire was administered online through Google Forms, utilizing a five-point Likert scale. The instrument used in this study underwent expert review and subsequent revisions. Afterward, the V coefficient (Aiken, 1985) was determined. The average result of the content validity test, based on the V coefficient, yielded an overall test value of 0.94, which is considered excellent (Putranta & Supahar, 2019). Once the content validity was established, the instrument was further field-tested. In this study, two approaches were applied to analyze the data. The first was the traditional ordinal approach, where the original Likert-scale responses were analyzed in their raw form. Each category was treated as ordered but not equally spaced, allowing the analysis to preserve the natural ordinal characteristics of the data. This approach reflects the common practice of using untransformed Likert responses to represent students' levels of agreement, providing a baseline for comparison with transformed data. The second approach involved transforming the same data using the summated rating scale method. The scaling process applied to the data from the field trial was based on the summated rating scale approach. Once scaled, psychometric characteristics of the instrument were analyzed, comparing both ordinal data obtained without scaling and scaled using a Summated Rating Scale (SRS) approach.

The instrument validation procedures were conducted to ensure that the measurement tool used is both valid and reliable. This process involved expert judgment and statistical analyses to evaluate the alignment of each item with the intended construct. Validation was carried out to guarantee that the collected data accurately reflect the quality and credibility of the measurement. The stages in instrument validation can be found in Figure 1.


Aiken's V
Confirmatory Factor Analysis
Partial Credit Model (PCM)
Instrument
Content Validity
Construct Validity
Item Response Theory
4 Reter
Loading Factor ≥ 0,3
Assumptions of Item Response Theory and Item Test Characteristics Analysis with IRT

Sources: Personal data (2024).

Figure 1. Stages of Instrument Validation Procedures

Item fit was evaluated using the PV_Q1 statistic, the use of PV_Q1 in this context serves to evaluate the appropriateness of data fit for the instrument, aligning with the total number of items included (Chalmers & Ng, 2017), to determine the compatibility of each item with the specified IRT models Partial Credit Model (PCM), Generalized Partial Credit Model (GPCM), and Graded Response Model (GRM). The PV_Q1 statistic assesses item-level fit by comparing observed and expected response patterns. Items with PV_Q1 values equal to or greater than 0.05 are considered to fit well within the model, indicating that the item responses align adequately with the assumptions and structure of the selected IRT model.

This study employs Item Response Theory (IRT) to compare different scaling approaches in assessing critical thinking instruments, following a structured methodology to ensure rigorous psychometric evaluation. Initially, the Summated Rating Scale (SRS) method is applied to transform ordinal Likert scale data into a continuous scale, thereby aligning with the assumptions required for advanced psychometric techniques. The analysis proceeds with the application of a polytomous IRT model to the original categorical data, enabling the evaluation of item characteristics and respondent abilities based on discrete response categories. Subsequently, a continuous response model (CRM) IRT is implemented using the SRS-transformed data to assess its psychometric properties under an alternative scaling approach. A comprehensive comparison between the polytomous and continuous IRT models is then conducted across four key dimensions: unidimensionality, invariance, item parameters, and ability estimation. First,Unidimensionality ensures that both models accurately measure a single latent trait, with the CRM potentially capturing finer nuances due to its continuous nature. Second, Invariance assesses the consistency of item and ability estimates across different subgroups or conditions, which is essential for the generalizability of the instrument. Comparing these aspects between the two models highlights whether the transformation from ordinal to continuous responses influences the stability of the psychometric properties and identifies which approach yields more precise and reliable estimates of critical thinking abilities (Payan-Carreira et al., 2022).

Third, item parameter estimation (difficulty, discrimination, and threshold parameters) determines how well items function within each model. The polytomous IRT model estimates thresholds for categorical responses, while the CRM IRT model estimates continuous response parameters. Differences in item parameter estimation may indicate whether scaling influences item functioning and interpretation. Finally, ability estimation assesses whether the latent trait estimates derived from both models align. Since the CRM IRT model converts Likert-scale data into a continuous metric, it may produce more precise ability estimates, but differences with the polytomous IRT model must be evaluated to determine practical implications. By systematically comparing these four aspects, researchers can identify which model provides a more accurate, reliable, and generalizable measurement of critical thinking skills, ensuring a scientifically sound assessment framework.

A comparative analysis of parameter estimates was conducted to evaluate differences in ability parameters generated by two scaling models. This analysis utilized statistical software that supports Item Response Theory (IRT), specifically through the use of the mirt package in R for polytomous IRT modeling (Chalmers, 2012), the EstCRM package for analyzing the Continuous Response Model (CRM) within the IRT framework (Zopluoglu, 2022), and the lavaan package for conducting Confirmatory Factor Analysis (CFA) in R (Rosseel, 2012). For data visualization and graphical representations, the ggplot2 package developed by Wickham, (2016) was employed.

Results

This study conducted a comprehensive analysis of instrument validity alongside an investigation of the differences using polytomous IRT based on data derived from measurement instruments without scaling, and continuous response model (CRM) IRT based on data that had been scaled using the summated rating approach. This study covers important aspects of IRT, including dimensionality assessment, model fit evaluation through empirical plots, item fit diagnostics, and parameter invariance tests. In addition, it also examines parameter estimation. Then, it provides an in-depth comparison of the two procedures to ensure a valid and meaningful comparison between the two approaches.

Instrument Content validity & Construct Validity

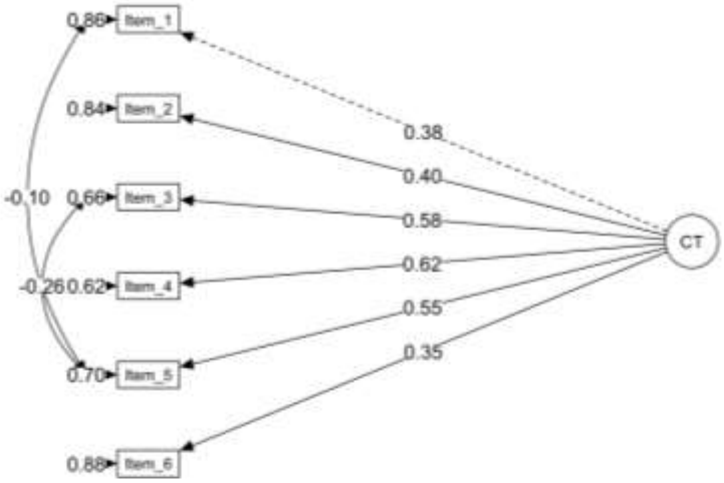
Instrument validity serves as a fundamental foundation to ensure that a measurement tool accurately assesses the intended construct. To evaluate content validity, this study employed Aiken's V formula (Aiken, 1985), involving four expert raters who assessed each item based on its relevance to the construct being measured. Aiken's V was chosen as it is a widely recommended quantitative approach for evaluating the consistency of expert judgments regarding the content appropriateness of an instrument. The V value is calculated by subtracting the lowest possible category score (c–1) from the total score given, and then dividing it by the maximum possible score range (n(c–1)).

Table 3. The Result of Aiken's V

Code	Rater				Σ	c-1	n(c-1)	V =Σ/n(c-1)	Criteria
Item_1	4	4	3	4	11	3	12	0.92	Valid
Item_2	4	4	4	3	11	3	12	0.92	Valid
Item_3	4	4	4	3	11	3	12	0.92	Valid
Item_4	4	4	4	4	12	3	12	1.00	Valid
Item_5	4	4	4	4	12	3	12	1.00	Valid
Item_6	4	4	4	3	11	3	12	0.92	Valid
Total								0.94	Valid

Sources: Personal data (2024).

Based on the results of the Aiken's V calculations presented in Table 3, all items were found to be content-valid, with V values ranging from 0.92 to 1.00. The overall Aiken's V score of 0.94 indicates a high level of agreement among the experts regarding the relevance of each item, demonstrating that the instrument possesses excellent content quality and is suitable for subsequent analytical stages. To assess construct validity, a Confirmatory Factor Analysis (CFA) was conducted on six items developed to measure the Critical Thinking (CT) construct. CFA enables researchers to confirm the theoretically hypothesized factor structure and evaluate the extent to which the empirical data support the measurement model (Bean & Bowen, 2021; Widaman & Revelle, 2023).



Sources: Personal data (2024).

Figure 2. Standardized Solution Diagram of First-order CFA Results

Figure 2 illustrates the path relationships between the Critical Thinking (CT) construct and its six associated items. The standardized path coefficients range from 0.35 to 0.62. According to Hair et al. (2019), for a sample size of approximately 250, a factor loading of 0.35 is considered acceptable. The results presented indicate that each item makes a significant contribution to the construct. Modification indices were applied to account for correlated error terms between Item_1 and Item_5, as well as between Item_3 and Item_5, reflecting similarities in content or formulation between these items. The evaluation results of various model fit indices are presented in the following table.

Table 4. Model Fit Indices

Fit Indices	Fit Criteria	Result	Criteria
p-value	≥ 0.05	0.14	Fit
Root Mean Square Error of Aproximation (RMSEA)	≤ 0.08	0.05	Fit
Goodness-of Fit Index (GFI)	≥ 0.90	0.99	Fit
Standardized Root Mean Residual (SRMR)	≤ 0.08	0.04	Fit
Normed Fit Index (NFI)	≥ 0.90	0.92	Fit
Tucker Lewis Index (TLI)	≥ 0.90	0.94	Fit
Comparative Fit Index (CFI)	≥ 0.90	0.97	Fit
Adjusted Goodness of Fit Indeks (AGFI)	≥ 0.90	0.96	Fit

Sources: Personal data (2024).

The fit indices presented in Table 4, follow the rules of thumb for fixed fit index cutoffs (Goretzko et al., 2024; Hair et al., 2019). All model fit index values indicate that the measurement model demonstrates an excellent fit with the empirical data, confirming a high level of model adequacy. Consequently, the Critical Thinking construct can be considered structurally validated.

Summated Rating Scale

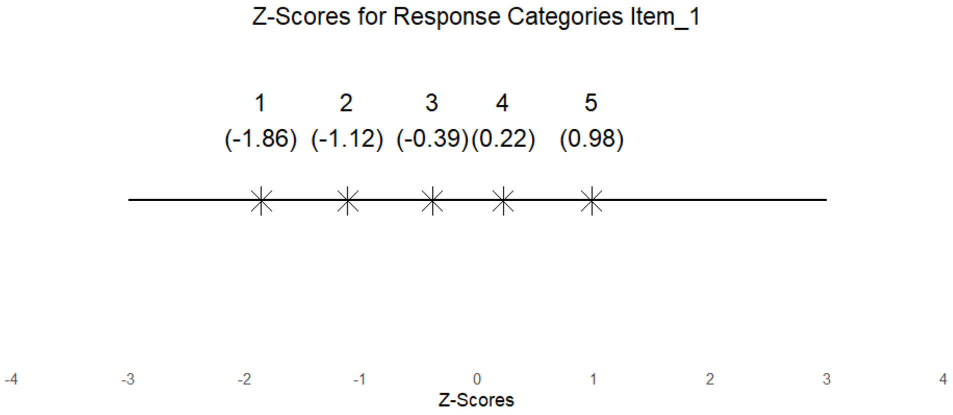
The scaling process for Likert-type instruments was conducted using the summated rating method as described by (Febriana & Setiawati, 2024). This approach involves aggregating respondents' answers into numerical scores, which are then summed to produce a total score for each individual. This method facilitates more precise measurement of constructs that are not directly observable, such as attitudes or perceptions. For instance, a sample item from this instrument is presented in the table below, illustrating the step-by-step calculation of summated ratings based on respondents' answers. This technique is widely applied in educational and psychological research to assess levels of agreement or disagreement with various statements. A detailed example of summated rating calculations for a specific item can be found in Table 5.

Table 5. Calculation of Summated Rating Scaling on One Item

No. Items	Category	F	Prop	Cum	Density	z	Scale
Item_1	1	17	0.06	0.06	0.03	-1.86	0.00
	2	36	0.13	0.20	0.13	-1.12	0.73
	3	80	0.30	0.50	0.35	-0.39	1.46
	4	48	0.18	0.68	0.59	0.22	2.07
	5	87	0.32	1.00	0.84	0.98	2.84

Sources: Personal data (2024).

The summated rating scaling process yielded z-scores for each response on every item. These outcomes demonstrate that the scaling method generates response scores that differ from those obtained without scaling. The scaled scores reveal that the intervals between responses on each item are not uniform or consistently equal to 1. For instance, in item_1, the score for category 1 shifts to -1.86, category 2 to -1.12, category 3 to -0.39, category 4 to 0.22, and category 5 to 0.98. If the lowest score is adjusted to 0, category 2 changes to 0.73, category 3 to 1.46, category 4 to 2.07, and category 5 to 2.84. This scaling approach produces a score distribution that more accurately captures the variations in response proportions across categories. Consequently, summated rating scaling offers a more proportional representation of the data distribution compared to the unscaled method, where category intervals are presumed equal. This refinement enhances measurement precision, particularly for advanced analyses such as construct validity assessments or factor analysis. The complete list of items transformed into summated ratings can be found in the appendix section.



Sources: Personal data (2024).

Figure 3. Distribution of Z Values in Response Categories: A Summated Rating Scale Approach

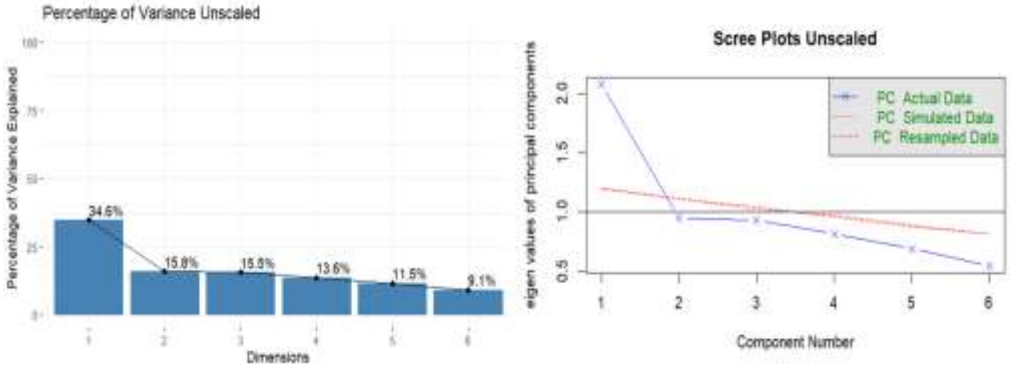
As depicted in Figure 3, the original z-scores for the response categories revealed substantial disparities: the lowest category (Category 1) registered a z-score of -1.86, while the highest (Category 5) scored 0.98. This non-uniform distribution underscores the inadequacy of treating ordinal categories as equally spaced units. Following normalization where the lowest category was anchored to zero the recalibrated scale exhibited proportional intervals: Category 2 (0.73), Category 3 (1.46), Category 4 (2.07), and Category 5 (2.84). The marked differences between adjacent categories (e.g., a 0.73 gap between Categories 1 and 2 versus 0.77 between Categories 4 and 5) highlight the inherent heterogeneity in response probabilities, which discrete ordinal models fail to capture.

Polytomous IRT

The quality of test items is not only shown by validity and reliability but can also be evaluated using Item Response Theory (IRT), which enables the separation of item characteristics from test-taker abilities through parameter invariance ensuring item properties remain consistent across different ability levels. Additionally, IRT upholds the principle of local independence, meaning that a test-taker's response to one item does not influence their responses to others. Furthermore, the assumption of unidimensionality ensures that each item measures only a single underlying ability or trait (Hambleton et al., 1991). Within IRT, several key parameters include difficulty (threshold), indicating how challenging an item is; discrimination, showing how well an item differentiates between test-takers of varying abilities; and distractors, which are incorrect options designed to reduce random guessing.

Dimensionality Assumption Test

The first critical assumption in psychometric analysis is unidimensionality, which posits that all items within a test measure a single latent trait or construct (Hartono et al., 2022). Fulfillment of this assumption strengthens the validity of subsequent analyses, as it ensures that item responses reflect a coherent underlying dimension. According to established criteria, unidimensionality is supported if the raw variance explained by the primary factor exceeds 20%, with values above 40% indicating strong unidimensionality and those surpassing 60% reflecting exceptional measurement precision (Hartono et al., 2022).



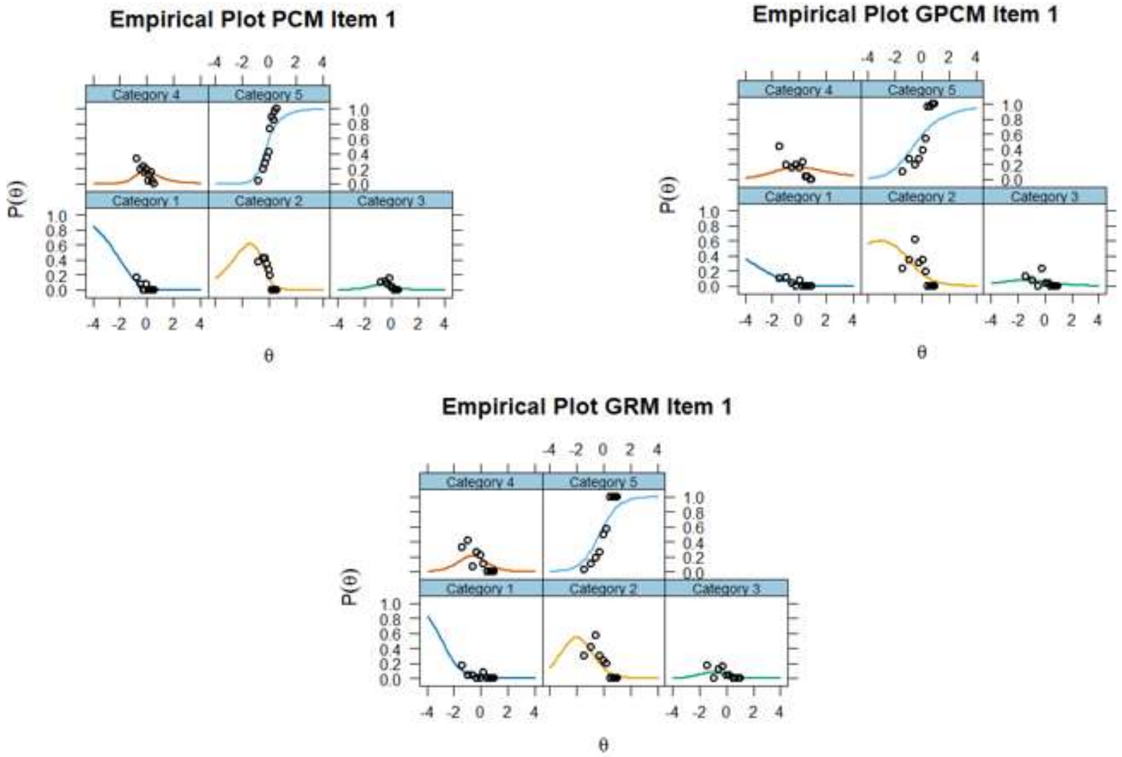
Sources: Personal data (2024).

Figure 4. Percentage of Variance and Scree Plot

As illustrated in Figure 4, the scree plot and eigenvalues derived from principal component analysis (PCA) reveal distinct dimensionality patterns. The first principal component (PC1) exhibits an eigenvalue of 2.08, accounting for 34.6% of the total variance significantly higher than subsequent components, all of which yield eigenvalues below 1.0. This sharp drop after the first component supports unidimensionality. However, the explained variance (34.6%) falls short of the 40% threshold for "strong" unidimensionality, instead positioning the instrument within the "moderate" range (Sumintono & Widhiarso, 2014). While this meets the minimum criterion (≥20%), the marginal variance explained implies that the unidimensional structure is present but not robustly optimized.

Model Fit Test Using Empirical Plot

In this section, a model fit test using an empirical plot was conducted to assess how well the model aligns with the observed data. This visual approach helps compare actual data distribution with model predictions, offering insight into the model's validity for this study.



Sources: Personal data (2024).

Figure 5. Empirical Plot Insights for Ordinal Scale Item_1

Figure 5 presents an empirical plot illustrating various IRT models. The plots display the probabilities of items analyzed using ordinal data. These visualizations capture the relationship between latent ability (θ) and the probability of selecting different categories across items for three primary models: the Partial Credit Model (PCM), the Generalized Partial Credit Model (GPCM), and the Graded Response Model (GRM). The points on each curve represent observed data from individuals' responses across different ability levels (θ). Each point corresponds to the actual selection of a specific category by respondents, mapped to their respective ability levels. In the context of IRT analysis, these points provide valuable insights into the frequency of category selection at varying ability levels, as well as the distribution and preference for specific categories.

The plots compare the performance of PCM, GPCM, and GRM on ordinal items. They illustrate how the probability of selecting each category (e.g., Category 1, Category 2, etc.) changes with shifts in ability (θ). Each model demonstrates distinct patterns: PCM shows sharp transitions between categories, GPCM exhibits smoother changes, and GRM captures more gradual variations. where responses are grouped into fewer categories.

For PCM, each point on the curve indicates the proportion of respondents selecting a specific category along the ability scale. For example, lower categories (e.g., Category 1 or 2) have points clustered at lower ability levels, reflecting a tendency for individuals with lower ability to choose these categories. Conversely, points for higher categories appear at higher ability levels, indicating that individuals with greater ability are more likely to select them. This pattern demonstrates how the probability of selecting a category increases as an individual's ability improves.

In GPCM and GRM, the points represent category selection based on individual ability levels but with notable differences in the behavior of category curves. For GPCM, higher-category points may exhibit broader or more dispersed curves, indicating greater variability in category selection among individuals with higher abilities. GRM, on the other hand, often displays smoother and more intricate transitions between categories, reflecting more dynamic changes in category preferences as ability evolves. PCM, while simpler, highlights sharper transitions compared to the other models.

Item Fit

Item fit evaluation plays a crucial role in ensuring the quality and validity of educational assessments. By identifying misfitting items, the measurement process becomes more trustworthy and precise. Table 6 presents item fit indices across the three measurement models tested earlier. Using the PV_Q1 statistic with a threshold of ≥ 0.05 (Chalmers & Ng, 2017). Items with PV_Q1 values below this threshold are flagged as "Not Fit," while those meeting or exceeding the threshold are labeled as "Fit." This evaluation helps identify items that may require revision or removal to improve the overall quality of the measurement instrument.

Table 6. Item Fit

Item	PCM	GPCM	GRM
	PV_Q1 (Fit ≥ 0.05)	**PV_Q1 (Fit ≥ 0.05)**	**PV_Q1 (Fit ≥ 0.05)**
Item_1	0.29 (Fit)	0.45 (Fit)	0.33 (Fit)
Item_2	0.00 (Not Fit)	0.01 (Not Fit)	0.03 (Not Fit)
Item_3	0.19 (Fit)	0.01 (Not Fit)	0.43 (Fit)
Item_4	0.45 (Fit)	0.15 (Fit)	0.08 (Fit)
Item_5	0.82 (Fit)	0.87 (Fit)	0.84 (Fit)
Item_6	0.59 (Fit)	0.36 (Fit)	0.54 (Fit)

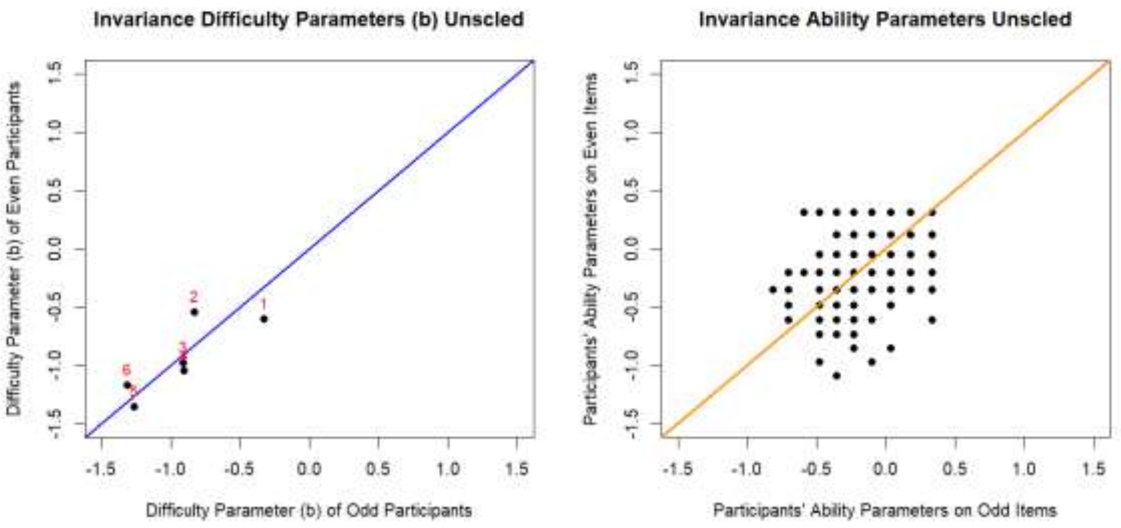
Sources: Personal data (2024).

In summary, Table 6 clearly demonstrates the performance of each item across the three models. Notably, the PCM consistently shows a strong fit for the majority of items, with only Item_2 failing to meet the fit criterion. In contrast, the GPCM and GRM exhibit inconsistencies, particularly with Item_3, where the GPCM fails to fit while the PCM and GRM both achieve a good fit. This inconsistency in the GPCM raises concerns about its reliability for certain types of data. Furthermore, the PCM's ability to maintain a good fit across multiple items (Item_1, Item_3, Item_4, Item_5, and Item_6) highlights its robustness and adaptability. The PCM's simplicity is one of its greatest strengths. Unlike the GPCM and GRM, which introduce additional parameters and complexity, the PCM provides a straightforward yet powerful framework for analyzing polytomous data. This simplicity not only makes the PCM easier to interpret but also reduces the risk of overfitting, ensuring that the model remains generalizable to new data. Moreover, the PCM's strong performance in this analysis underscores its suitability for a wide range of applications, from educational testing to psychological assessments.

Invariance parameters

The comparison of invariance parameters in IRT is crucial to ensure item parameters particularly difficulty (b) and ability (θ) remain consistent across groups, supporting the fairness and validity of the instrument. In this study, only the difficulty and ability parameters are analyzed, as the Partial Credit Model (PCM) was selected during model fit testing, which fixes the discriminant parameter (a) to 1 for all items. The decision to exclude the discriminant parameter (a) from this analysis is driven by the results of the initial model fit test, which identified the PCM as the most appropriate model for this data.

The invariance of item parameters across groups or conditions is crucial to ensuring that the measurement model operates similarly regardless of external factors, thus reinforcing the fairness and comparability of the assessment outcomes. By examining the difficulty (b) and ability (θ) parameters, the study seeks to evaluate whether these key dimensions of IRT remain stable and unbiased when applied to diverse populations or test conditions. This focus is essential, as inconsistencies in these parameters can distort the interpretation of test results, potentially leading to invalid conclusions about a student's proficiency or the effectiveness of the measurement tool. Under this model, the value of the discriminant parameter is fixed at 1 for all items, rendering it irrelevant for the comparison of invariance parameters. To provide a clearer understanding, the differences in these invariance parameters will be illustrated in Figure 6.



Sources: Personal data (2024).

Figure 6. Invariance Difficulty (b) and Ability Parameters Generated from Unscaled Data

Figure 6 visually compares difficulty (b) and ability (θ) parameters across groups. The difficulty parameter (b) across different groups or conditions. The results indicate a high correlation of 0.84 for the difficulty parameter, suggesting that item difficulty remains consistently stable across the diverse groups examined. This high level of correlation indicates consistent item difficulty, ensuring the assessment is fair and reliable for all participants regardless of group. Such consistency in item difficulty is critical for maintaining the overall fairness of the measurement process.

Conversely, the ability parameter (θ) reveals a lower correlation of 0.37, indicating greater variability in the estimates of participants' abilities across the different items. This variability in the ability parameter suggests that the estimates of ability may differ more significantly across participants, possibly due to the varying characteristics of the items themselves or the diverse nature of the participant pool. Such variability in ability estimates can reflect differences in how participants engage with the items, highlighting the need for a more nuanced interpretation of the ability parameter across different groups or conditions. This variation is an important consideration for researchers and practitioners using IRT models, as it points to the complexity of measuring and interpreting student abilities in diverse contexts.

Parameter Estimation

The analysis of parameter estimation in the context of IRT polytomus provides profound insights into how data transformation influences the estimation of item parameters, under the PCM. The comparison, as illustrated in Table 7, reveals significant variations in parameter estimates across the two data types, underscoring the critical role of scaling in the interpretation and accuracy of IRT models.

Table 7. Analysis of Parameter Estimation: Unscaled and Scaled Comparison

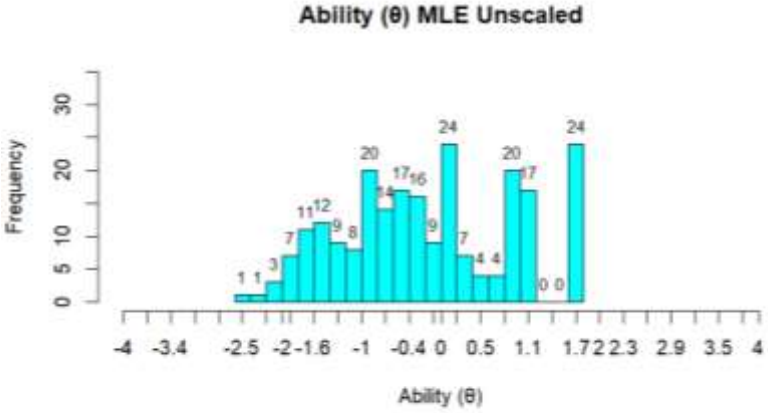
Item	Unscaled					
	a	b1	b2	b3	b4	location
Item_1	1	-1.19	-1.04	0.49	-0.39	-0.46
Item_2	1	-2.30	1.02	-1.15	-1.26	-0.66
Item_3	1	-3.32	0.67	-1.86	-1.19	-1.08
Item_4	1	-2.79	0.61	-1.54	-1.79	-1.11
Item_5	1	-1.61	-0.85	0.95	-4.14	-1.29
Item_6	1	-1.55	-0.20	-1.54	-1.92	-1.23

Sources: Personal data (2024).

The comparison between unscaled and scaled parameter estimation in Table 7, presents a detailed comparison of parameter estimates for eight items, showcasing the stark differences between unscaled and scaled data. A notable distinction lies in the location parameter, a key distinction lies in the location indices parameter, also referred to as generalized item difficulty is calculated as the average of accumulated thresholds. In the ordinal model, this parameter is computed as the average of the accumulated thresholds (b1, b2, b3, ...) (Ali et al., 2015). In the ordinal model, this parameter is computed as the average of the accumulated thresholds (b1, b2, b3, etc.), reflecting the relative positioning of items on the latent trait continuum. The discrepancies observed between unscaled and scaled data indicate that the transformation process affects not only individual threshold estimates but also the overall difficulty structure of the items.

Maximum Likelihood Estimation (MLE) of Ability (θ) Unscaled

Maximum Likelihood Estimation (MLE) is a technique in statistics that aims to estimate model parameters by maximizing the likelihood function. This function evaluates how well the parameters account for the observed data. When applied to ability estimation, MLE helps identify the most probable ability level (θ) for individuals, derived from their performance on test items. The objective is to determine the θ value that best aligns with the observed data.



Sources: Personal data (2024).

Figure 7. Ability Distribution of Respondents in an IRT Model (MLE Unscaled)

Figure 7 illustrates the distribution of ability (θ) as estimated using Maximum Likelihood Estimation. The ability values range from -4 to 4, with certain values such as -2, -1, 0.5, and 1.7 showing higher frequencies or likelihoods. This suggests that these ability levels are more common or more likely given the data. The graph provides a visual representation of how ability is distributed across the sampled population, highlighting the most probable ability levels. This type of analysis is crucial in fields like psychometrics and educational testing, where understanding the distribution of abilities can inform test design and interpretation.

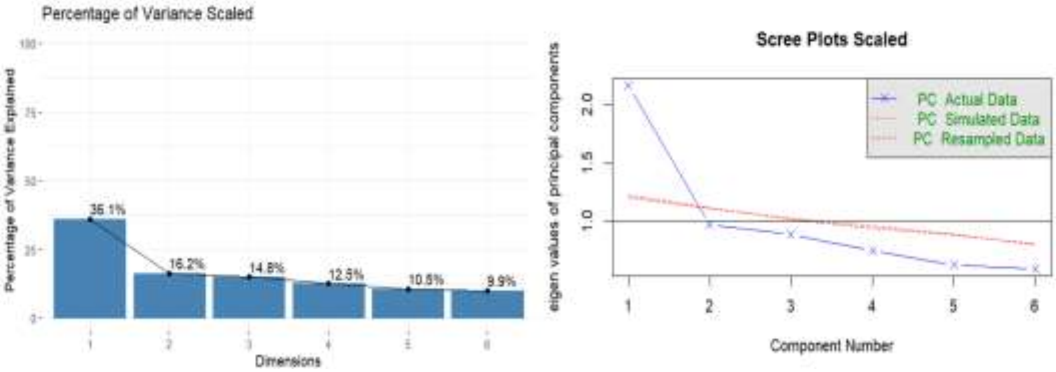
The unscaled nature of the θ values suggests that the data is presented in its raw form, without normalization, which can be important for understanding the original distribution and variability in the data. MLE encounters a notable limitation when applied to the total sample of 268 respondents. Specifically, while the ability estimates for 228 respondents are successfully obtained, the remaining 40 respondents are assigned infinite ability values or are deemed unidentifiable in terms of ability estimation. This phenomenon occurs because MLE relies on the variability of response patterns to derive ability estimates. When respondents answer all items correctly or incorrectly, the likelihood function does not provide a finite maximum, leading to an indeterminate ability estimate.

Continuous Response Model (CRM) IRT

Since the objective is to compare scaled data with non-scaled data, the IRT analysis must also employ an appropriate model that accommodates continuous responses. In this context, Samejima's Continuous Response Model (CRM) (Samejima, 1973), is particularly suitable, as it allows for the modeling of test items where responses are not limited to discrete categories but instead exist on a continuous scale. This approach provides a more nuanced understanding of item characteristics and test-taker abilities, ensuring that the analysis remains robust and theoretically sound. By leveraging CRM, researchers can better capture the subtleties of item functioning in assessments that involve gradations of correctness rather than binary or ordinal responses. Consequently, this model enhances the precision of item parameter estimation while maintaining the core assumptions of IRT, including unidimensionality and parameter invariance.

Dimensionality Assumption Test

Ensuring unidimensionality in CRM within IRT is crucial, as it guarantees that all items collectively measure a single latent trait, preserving the accuracy of parameter estimation and score interpretation. Since CRM deals with responses on a continuous scale, such as those derived from the SRS approach, any deviation from unidimensionality can result in biased ability estimates and misinterpretations of test outcomes. The results of the unidimensionality test are presented in Figure 8.



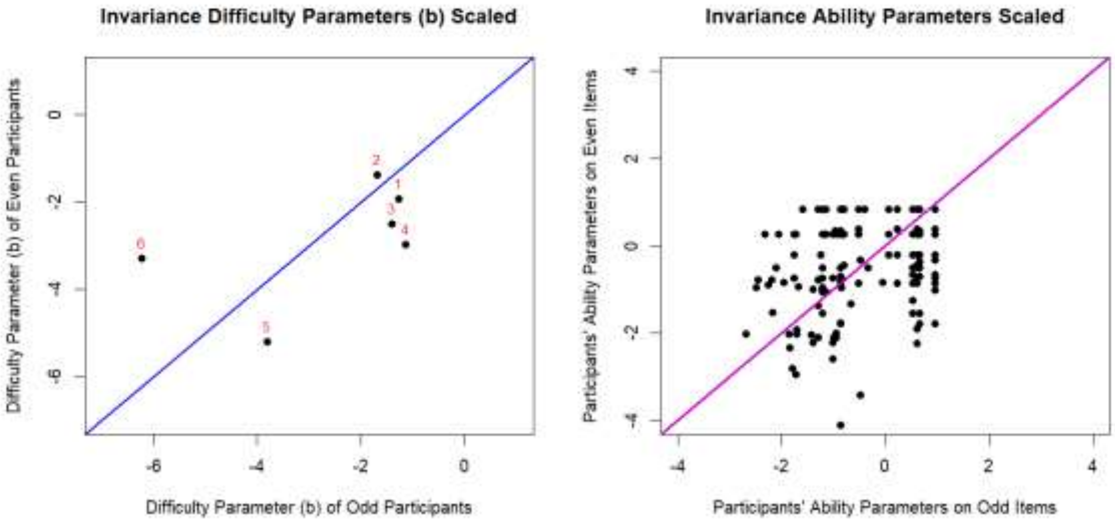
Sources: Personal data (2024).

Figure 8. Percentage of Variance and Scree Plot

The scree plot shows a sharp drop in eigenvalues after the first component, which exceeds 2, while the others fall near or below 1, indicating a dominant single dimension and supporting unidimensionality. According to established criteria, a factor explaining at least 20% of the variance supports unidimensionality, though a stronger threshold of 40% or more is preferred. In this case, the first component accounts for 36.1% of the variance, while the second contributes 16.2%, bringing the cumulative variance of the first two dimensions to 52.3%. level indicative of a reasonably strong primary dimension.

Invariance parameters

Item parameter invariance in Continuous Response IRT ensures that item characteristics remain stable regardless of differences in respondents' ability levels, supporting measurement validity. Likewise, ability parameter invariance confirms that proficiency estimates reflect true ability, unaffected by item difficulty. In this study, respondents were split into odd- and even-numbered groups to test the consistency of item parameters across samples. This analysis aims to verify the model's fairness and reliability by confirming parameter stability across respondent groups. The results of the invariance difficulty (b) and ability parameters generated from scaled data are presented in Figure 9.



Sources: Personal data (2024).

Figure 9. Invariance Difficulty (b) and Ability Parameters Generated from Scaled Data

Figure 9 illustrates the relationship between item difficulty parameters (b), which reflects the consistency of item calibration across groups. Ideally, if the difficulty parameters were perfectly invariant, all data points would align along the identity line. With a correlation coefficient of 0.53, the results suggest a moderate level of stability in item difficulty estimates. However, Item_5 and Item_6 exhibit noticeable deviations, indicating potential inconsistencies. These discrepancies may arise from differential response patterns across groups, item bias, or other latent factors affecting measurement precision. While the overall model performance remains satisfactory, further investigation into these specific items is warranted to enhance the instrument's robustness. The invariance analysis of ability parameters examines the consistency of respondent ability estimates, where identical estimates are expected if no significant differences exist between item sets. However, with a correlation of 0.46, the relationship is relatively weak to moderate, suggesting potential inconsistencies in ability estimation.

Parameter Estimation

The analysis of parameter estimation CRM in IRT extends traditional dichotomous and polytomous models by accommodating responses that exist on a continuous scale, rather than discrete categories. Unlike polytomous models, such as the PCM or the GPCM, which define ordered response categories, the CRM assumes that responses are continuous values, making it particularly suited for assessments where answers are measured on an interval or ratio scale, such as time taken to complete a task, reaction times, or ratings on an analog scale. The results of the parameter estimation generated from scaled data are presented in Table 8.

Table 8. Parameter Estimation Generated from Scaled Data

Item	**scaled**		
	a	**b**	**alpha (α)**
Item_1	0.27	-1.85	0.84
Item_2	0.88	-1.43	2.09
Item_3	0.65	-2.44	1.53
Item_4	0.77	-2.47	1.65
Item_5	0.51	-4.54	1.04
Item_6	0.37	-4.38	0.95

Sources: Personal data (2024).

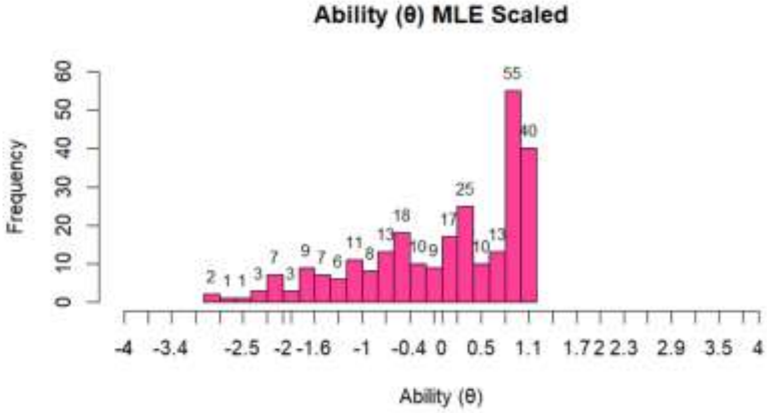
In CRM, the item parameters include discrimination (a), difficulty (b), and an additional scaling parameter (α), which captures the variability in continuous responses. The discrimination parameter (a) functions similarly to that in dichotomous IRT models, indicating the sensitivity of an item in differentiating between individuals with different levels of the latent trait. The difficulty parameter (b) represents the point on the latent trait continuum where the probability density of responses is centered. The additional scaling parameter (α) controls the distance between item characteristic curves, which can affect the shape and separability of the curves for each item. The larger the value of α, the higher the flexibility of the scale that can affect the sensitivity of changes in respondent ability to the probability of answering correctly.

As illustrated in Table 8, Item_1 has parameters a = 0.27, b = -1.85, and $\alpha = 0.84$. The relatively low value of a indicates that this item is less effective in distinguishing participants with different abilities. The value of b = -1.85 indicates that this item is quite easy to answer correctly by low ability participants. Meanwhile, the value of $\alpha = 0.84$ indicates that the distance between the item characteristic curves is relatively small, so the change in probability of correct answer occurs more gradually than items with a larger α.

By considering these three parameters, item analysis can be done more comprehensively, allowing a deeper understanding of the characteristics of each item in a test. Parameter estimates for scaled data exhibit distinct variations in discrimination and response dispersion, highlighting the nuanced way continuous responses are modeled compared to traditional categorical IRT frameworks. These distinctions make CRM particularly valuable in educational and psychological assessments where responses extend beyond simple categorical selections.

Maximum Likelihood Estimation (MLE) of Ability (θ) Scaled

Maximum Likelihood Estimation (MLE) identifies the most likely ability level (θ) for each individual based on their response patterns, aligning with the probabilistic structure of IRT. While MLE is effective for estimating ability, it can be unreliable in cases of extreme scores such as all correct or all incorrect responses where estimates may become undefined. However, in this study, using a summated rating scale analyzed under a continuous IRT model, no infinite θ values were found. This contrasts with dichotomous IRT models, where extreme patterns often lead to indeterminate ability estimates. As shown in Figure 10.



Sources: Personal data (2024).

Figure 10. Ability Distribution of Respondents in an IRT Model (MLE Scaled)

Figure 10 presents a graphical representation of the MLE for ability (θ) under a continuous response IRT model. Unlike polytomous models, which handle discrete response categories, the continuous response model treats ability as a fluid, non-categorical construct, allowing for a more nuanced representation of individual performance. The distribution reveals that ability levels such as -2, -1, 0.5, and 0.9 exhibit higher likelihoods, suggesting that these values are more prevalent within the assessed population. This visualization offers critical insights into the underlying structure of ability estimates, aiding in the refinement of measurement precision and model assumptions. Additionally, the unscaled nature of the θ values ensures that raw measurement properties remain intact, facilitating deeper analysis of individual differences without distortion from normalization techniques. This approach is particularly valuable in educational and psychological assessments where fine-grained differentiation of abilities is essential for valid interpretations.

Comparative Analysis of Item Response Theory Unscaled and Scaled Data

This study offers a detailed comparison between Polytomous IRT, based on data from critical thinking instruments, and CRM IRT, which utilizes data scaled through SR, focusing on their distinct methodological frameworks, theoretical foundations, and practical applications. While both models assume unidimensionality meaning all items measure a single latent trait CRM IRT shows a stronger fit to this assumption, as indicated by its scree plot where the first principal component accounts for 36.1% of variance versus 34.6% for Polytomous IRT. This difference stems from the nature of the data each model processes: Polytomous IRT handles ordinal responses assuming equal intervals between categories, potentially distorting construct dimensionality, whereas CRM IRT works with continuous interval-scale data that allow for nonuniform intervals, better capturing proportional response differences and minimizing secondary factor interference, thereby enhancing the representation of the primary latent trait. The analysis highlights their respective strengths and limitations across key aspects like dimensionality, model fit, parameter invariance, and measurement precision, providing insights into their suitability for various psychometric contexts.

The evaluation of parameter invariance revealed fundamental distinctions between the polytomous IRT and the CRM IRT frameworks, particularly in the stability of difficulty parameters (b) and ability estimates (θ). Polytomous IRT exhibited a strong correlation in difficulty parameters ($r = 0.84$), indicating high stability across different test conditions. However, the variability in ability estimates ($r = 0.37$) suggests a degree of inconsistency, likely arising from the model's sensitivity to extreme response patterns. This phenomenon is often observed in polytomous frameworks, where the interpretation of extreme responses can be affected by individual response tendencies rather than true latent trait differences. The observed lower reliability in θ estimates implies that while item calibration remains robust, the estimation of examinee abilities may be more susceptible to variations in response behaviors, particularly in cases where respondents exhibit tendencies toward extreme or central categories.

Conversely, the CRM IRT model showed moderate invariance in item difficulty ($r = 0.53$) but greater stability in ability estimates ($r = 0.46$), reflecting a trade-off compared to polytomous IRT, which demonstrated higher item parameter stability but lower reliability in estimating abilities. The continuous scaling of CRM allows for a more fluid representation of item difficulty and better captures nuanced trait variations, especially in cases of extreme or perfect scores, though it may introduce variability based on sample distribution. These findings highlight key distinctions in how each model handles measurement invariance: polytomous IRT excels in item stability but may hinder accurate classification, while CRM IRT offers more consistent ability estimates, making it well-suited for assessments prioritizing detailed trait differentiation.

The comparison between the Partial Credit Model (PCM) and the Continuous Response Model (CRM) in Item Response Theory (IRT) highlights fundamental differences in how item parameters are estimated and interpreted. PCM, as a polytomous IRT model, is designed for ordinal data, capturing difficulty thresholds (b1, b2, b3, b4) that represent transitions between response categories. In this model, discrimination (a) is fixed at 1.0, which limits its flexibility in modeling varying item discriminations. In contrast, CRM IRT, as implemented in this analysis, applies a transformation to the same ordinal data using the summated rating scale method, treating the responses as a continuous variable. This transformation introduces additional parameters, including discrimination (a) and a scaling factor (α\alpha), which accounts for variations in response scaling.

The parameter estimation results further illustrate these distinctions. In PCM, the location parameter reflects the central tendency of difficulty thresholds, while in CRM, a single difficulty parameter (b) is estimated, capturing the item's position along the latent trait continuum. For example, Item_1 in PCM has multiple thresholds ranging from -1.19 to 0.49, whereas in CRM, it has a single difficulty estimate of -1.85. Additionally, the discrimination parameter in CRM varies across items, with Item_1 showing a lower aa value (0.27) compared to Item_2 (0.88), reflecting different sensitivities in distinguishing respondents' abilities. The scaling parameter (α\alpha) in CRM, such as 0.84 for Item_1, further refines response variability, a feature absent in PCM.

If the data used in CRM IRT is the same as in PCM but has been scaled using the summated rating method, the distinction between the two models lies in how the responses are processed rather than in the nature of the data itself. PCM models categorical transitions explicitly, making it suitable for ordinal data with distinct response levels. CRM IRT, however, assumes that the scaled ordinal responses approximate a continuous latent trait, allowing for greater flexibility in estimating parameters. This explains why discrimination (aa) is fixed in PCM but varies in CRM, and why CRM introduces the scaling parameter (α\alpha) to capture response spacing along the latent trait continuum.

Overall, the choice between PCM and CRM IRT depends on the intended analytical perspective rather than just the data type. PCM is advantageous for modeling ordered categorical responses where category transitions need to be explicitly defined, such as Likert-scale items. Meanwhile, CRM IRT provides a more flexible framework for analyzing ordinal data that has been transformed to approximate continuous measurement, making it useful when greater adaptability in discrimination and response scaling is required. Thus, while both models analyze the same underlying data, their differing assumptions and estimation approaches lead to distinct interpretations of item characteristics.

Estimating individuals' ability levels (θ) based on MLE is a fundamental approach in IRT. MLE applied to unscaled data directly estimates ability from raw responses, where item responses are typically dichotomous (e.g., correct or incorrect). This approach assumes that ability estimates maximize the likelihood of observed response patterns, but it encounters limitations when extreme response patterns occur specifically, when respondents answer all items correctly or incorrectly. In such cases, the likelihood function does not reach a finite maximum, resulting in infinite or undefined ability estimates. This limitation is particularly problematic in dichotomous IRT models, where the absence of variability in response patterns prevents MLE from deriving meaningful estimates for all respondents. As observed in the unscaled MLE analysis, 40 out of 268 respondents experienced such issues, highlighting a critical drawback of relying on raw response data in estimating ability.

In contrast, scaled MLE addresses the limitations of unscaled MLE by transforming ordinal responses into a continuous metric using a summated rating scale approach. This transformation enhances response variability and ensures finite, stable ability estimates, making it particularly suitable for polytomous or continuous IRT models. While traditional MLE is standard for dichotomous data, it is prone to infinite estimates when faced with extreme scores, often requiring Bayesian methods or constraints. In contrast, scaled MLE retains the strengths of likelihood maximization while providing a more robust and interpretable framework. This highlights the theoretical and practical value of data scaling techniques in psychometric modeling, especially in mitigating ceiling and floor effects.

Ultimately, the choice between unscaled and scaled MLE depends on the nature of the test data and the desired precision in ability estimation. While MLE in unscaled data is straightforward and commonly applied in traditional IRT models, it can be limited by extreme response behaviors. Scaled MLE, by contrast, provides a more stable estimation framework, particularly when dealing with polytomous response formats. As educational and psychological assessments increasingly adopt sophisticated measurement models, leveraging scaling techniques can enhance the reliability and validity of ability estimates, ensuring more meaningful interpretations of respondent performance.

Discussion

The findings of this study underscore significant distinctions between the impacts of using interval data generated through the Summated Rating Scale (SRS) transformations and ordinal data without scaling. analyzed via polytomous IRT and Continuous Response Model (CRM) IRT, respectively. The results align with prior research while addressing gaps in comparative methodologies, particularly in the context of critical thinking assessment. The results confirm the hypotheses outlined in the introduction, demonstrating that while both methods yield acceptable model fit for most items, the SRS offers a more nuanced representation of student abilities. This aligns with the limitations of ordinal scales, which often fail to maintain consistent intervals between response categories, thereby affecting the accuracy of statistical analyses (Casper et al., 2020). For instance, the non-uniform z-score intervals in SRS-scaled data (Figure 3) corroborate Kusmaryono et al. (2022) and Zou et al. (2023), who emphasized that ordinal-to-interval transformations enhance measurement precision by normalizing response distributions. This refinement is critical for advanced analyses like factor validity, as SRS-generated data better reflects proportional differences in response probabilities, reducing distortions inherent in ordinal assumptions.

The superiority of the Partial Credit Model (PCM) in item fit and parameter invariance echoes Hambleton et al., (1991), who highlighted PCM's robustness in handling polytomous data. While GPCM and GRM exhibited inconsistencies particularly for Item_3 PCM's simplicity and stability align with Robitzsch and Lüdtke, (2020), who argued that fewer parameters reduce overfitting risks. Similarly, the moderate unidimensionality (34.6% variance explained, Figure 4) resonates with Sumintono and Widhiarso (2014) criteria, suggesting that while the instrument measures a single latent trait, further optimization is needed for stronger dimensionality.

The invariance analysis revealed stark contrasts between scaled and unscaled data. Polytomous IRT showed strong difficulty parameter stability ($r = 0.84$) but inconsistent ability estimates ($r = 0.37$), reflecting results by Dai et al. (2021), who noted that ordinal models struggle with extreme response patterns. Conversely, CRM IRT's moderate difficulty invariance ($r = 0.53$) but improved ability stability ($r = 0.46$) aligns with Samejima (1973) assertion that continuous models mitigate ceiling/floor effects through probabilistic scaling. These findings address gaps identified by Shaw et al. (2020) and Tobón and Luna-nemecio (2021), who called for comparative studies on scaled vs. ordinal IRT applications.

Notably, SRS-scaled data resolved MLE's infinite ability estimates (Figure 10), a limitation prevalent in unscaled dichotomous models (Jebb et al., 2021). This aligns with Astuti et al. (2024), who advocated for scaling to enhance estimation robustness. Furthermore, the discriminative variability in CRM parameters (Table 8) supports Alamrani et al. (2023), who emphasized that continuous models better capture nuanced trait differences, crucial for complex constructs like critical thinking. These results underscore the need to transition toward SRS in contexts requiring precision, as advocated by Mustika et al. (2019) and Ebil et al. (2020). However, the retained utility of ordinal scales for stable difficulty calibration (Robitzsch, 2021) suggests a complementary approach. Educators must balance methodological rigor with practicality, as cautioned by (Kinel et al., 2021), ensuring assessments align with educational objectives.

The application of IRT in this study is pivotal in assessing the effectiveness of ordinal data and SR-scaled data in measuring critical thinking skills. IRT offers a more nuanced approach compared to Classical Test Theory (CTT), as it models the relationship between item responses and latent traits non-linearly, providing deeper insights into item characteristics and respondent abilities (Lord & Novick, 2008). The study's findings confirm that IRT is particularly effective in handling the complexities of both ordinal and continuous data, addressing the limitations of CTT, which primarily distinguishes between true scores and error scores without accounting for the probabilistic nature of item responses.

This study bridges the theoretical gap highlighted by Mohamadi (2018) and Lindner and Lindner (2024), offering empirical evidence on IRT-based comparisons. Future research should explore hybrid models and longitudinal applications to refine assessment frameworks, advancing vocational education's capacity to nurture critical thinking in dynamic environments. By integrating SRS and IRT, stakeholders can develop instruments that transcend ordinal limitations, fostering equitable and precise evaluations of student competencies.

Conclusion

This study provides a comprehensive comparative analysis of the psychometric properties of interval data derived from the Summated Rating Scale (SRS) and ordinal data analyzed through polytomous Item Response Theory (IRT) in assessing critical thinking skills. The findings yield significant insights into the methodological and practical implications of scaling techniques and measurement models, aligning with the research objectives to evaluate their distinct impacts on precision, validity, and reliability. First, the

SRS transformation demonstrated its efficacy in addressing the limitations of ordinal scales by converting categorical responses into interval-level data. This process resolved the inherent issue of unequal category intervals, as evidenced by the normalized z-score distributions (e.g., recalibrated scores for Categories 1–5: 0.00, 0.73, 1.46, 2.07, 2.84). Normalization improves measurement precision, enabling advanced statistical analyses that would be biased under ordinal assumptions. These results underscore the importance of scaling in contexts requiring detailed differentiation of latent traits, particularly for constructs like critical thinking, where proportional response differences reflect varying proficiency levels.

Second, the comparison of IRT frameworks revealed critical distinctions. While the Partial Credit Model (PCM) exhibited robust item fit and strong invariance in difficulty parameters (r = 0.84), its reliance on ordinal data led to instability in ability estimates (r = 0.37), particularly for respondents with extreme response patterns. In contrast, the Continuous Response Model (CRM) IRT, applied to SRS-scaled data, produced more stable ability estimates (r = 0.46) and eliminated infinite values in Maximum Likelihood Estimation (MLE). This underscores CRM's superiority in handling continuous data and mitigating ceiling/floor effects, thereby enhancing the interpretability of latent trait measurements. The implications of these findings extend to both theoretical and practical domains. For educators and assessment designers, adopting SRS scaling is imperative in high-stakes evaluations, as it ensures equitable interpretations of student competencies by accounting for non-uniform response distributions. The integration of CRM IRT further supports the development of adaptive assessments that accommodate diverse ability levels, a critical consideration in vocational education. This study bridges a gap in psychometric literature by empirically validating the benefits of combining scaling techniques with IRT, providing a replicable framework for future research on complex constructs.

However, the moderate unidimensionality observed in both models (34.6% variance explained for PCM vs. 36.1% for CRM) highlights the need for instrument refinement to strengthen the alignment of items with the latent trait. Future studies should explore hybrid models that leverage PCM's parameter stability and CRM's flexibility in continuous scaling. Longitudinal research could also investigate how scaled assessments influence long-term tracking of skill development, particularly in dynamic educational environments. In conclusion, this study advocates for a paradigm shift toward interval-based methodologies in psychometric analysis. By prioritizing SRS transformations and CRM IRT, stakeholders can advance the precision and fairness of critical thinking assessments, ensuring they meet the evolving demands of modern education. These advancements refine measurement practices and empower institutions to cultivate analytical skills crucial for vocational and lifelong success.

Acknowledgment

We extend our deepest gratitude to the Graduate School of Yogyakarta State University for the support and opportunities provided, so that this research can be carried out properly.

Conflict of Interest

The authors declare that there is no conflict of interest in the publication of this research. All data and findings are presented objectively without interference from any party. This research was purely conducted for academic purposes and the development of science.

Authors Contribution

AAM is fully responsible for the content of the article, starting from the introduction, method, data collecting, data analysis, and reporting research results. ASF played a role in the preparation of research instruments and data collecting. SH, NHPSP, and HR helped provided suggestions related to method used, data analysis, and improvements to the article manuscript.

References

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings, Educational and Psychological Measurumen, 45(1), 131–142. https://doi.org/10.1177/0013164485451012

Akour, I. A., Al-Maroof, R. S., Alfaisal, R., & Salloum, S. A. (2022). A conceptual framework for determining metaverse adoption in higher institutions of gulf area: An empirical study using hybrid SEM-ANN approach. Computers and Education: *Artificial Intelligence, 3*, 100052. https://doi.org/10.1016/j.caeai.2022.100052

Alamrani, S., Gardner, A., Falla, D., Russell, E., Rushton, A. B., & Heneghan, N. R. (2023). Content validity of the Scoliosis Research Society questionnaire (SRS-22r): A qualitative concept elicitation study. *PLoS ONE, 18*(5), 1–21. https://doi.org/10.1371/journal.pone.0285538

Ali, U. S., Chang, H., & Anderson, C. J. (2015). Location indices for ordinal polytomous items based on item response theory. *ETS Research Report Series*, *2015*(2), 1–13. https://doi.org/10.1002/ets2.12065

Alordiah, C. O., & Oji, J. (2024). Test equating in educational assessment: A comprehensive framework for promoting fairness, validity, and cross-cultural equity. *Asian Journal of Assessment in Teaching and Learning, 14*(1), 70–84. https://doi.org/10.37134/ajatel.vol14.1.7.2024

Astuti, N. D., Hajaroh, M., Prihatni, Y., Setiawan, A., Setiawati, F. A., & Retnawati, H. (2024). Comparison of KMO results, eigen value, reliability, and standard error of measurement: Original & rescaling through summated rating scaling. *Jurnal Pengukuran Psikologi dan Pendidikan Indonesia, 13*(2), 199–217. https://doi.org/10.15408/jp3i.v13i2.36684

Bean, G. J., & Bowen, N. K. (2021). Item response theory and confirmatory factor analysis: Complementary approaches for scale development. *Journal of Evidence-Based Social Work, 18*(6), 597–618. https://doi.org/10.1080/26408066.2021.1906813

Casper, W. C., Edwards, B. D., Wallace, J. C., Landis, R. S., & Fife, D. A. (2020). Selecting response anchors with equal intervals for summated rating scales. *Journal of Applied Psychology, 105*(4), 390–409. https://doi.org/10.1037/apl0000444

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6). https://doi.org/10.18637/jss.v048.i06

Chalmers, R. P., & Ng, V. (2017). Plausible-value imputation statistics for detecting item misfit. *Applied Psychological Measurement, 41*(5), 372–387. https://doi.org/10.1177/0146621617692079

Dai, S., Vo, T. T., Kehinde, O. J., He, H., Xue, Y., Demir, C., & Wang, X. (2021). Performance of polytomous irt models with rating scale data: An investigation over sample size, instrument length, and missing data. *Frontiers in Education, 6*, Article 721963, 1–18. https://doi.org/10.3389/feduc.2021.721963

Ebil, S. Hj., Salleh, S. M., & Shahrill, M. (2020). The use of E-portfolio for self-reflection to promote learning: A case of TVET students. *Education and Information Technologies, 25*(6), 5797–5814. https://doi.org/10.1007/s10639-020-10248-7

Febriana, B. W., & Setiawati, F. A. (2024). Increasing measurement accuracy: Scaling effect on academic resilience instrument using Method of Successive Interval (MSI) and Method of Summated Rating Scale (MSRS). *Jurnal Penelitian Dan Evaluasi Pendidikan, 28*(1), 32–42. https://doi.org/10.21831/pep.v28i1.69334

Fialho, L., & Zyngier, S. (2023). Quantitative methodological approaches to stylistics. In *The Routledge handbook of stylistics*. Routledge.

Goretzko, D., Siemund, K., & Sterner, P. (2024). Evaluating model fit of measurement models in confirmatory factor analysis. *Educational and Psychological Measurement, 84*(1), 123–144. https://doi.org/10.1177/00131644231163813

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis*. Cengage Learning.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.

Hartono, W., Hadi, S., Rosnawati, R., & Retnawati, H. (2022). Uji kecocokan model parameter logistik soal diagnosa kemampuan matematika dasar. *JNPM (Jurnal Nasional Pendidikan Matematika), 6*(1), 125. https://doi.org/10.33603/jnpm.v6i1.5899

Jebb, A. T., Ng, V., & Tay, L. (2021). A review of key likert scale development advances: 1995–2019. *Frontiers in Psychology, 12*, Article 637547, 1–14. https://doi.org/10.3389/fpsyg.2021.637547

Kadigi, R. M. J., Mgeni, C. P., Kangile, J. R., Aku, A. O. ati, & Kimaro, P. (2023). Can a legal game meat trade in Tanzania lead to reduced poaching? Perceptions of stakeholders in the wildlife industry. *Journal for Nature Conservation, 76*, 1–17. https://doi.org/10.1016/j.jnc.2023.126502

Kadim, A., & Sunardi, N. (2021). Financial management system (QRIS) based on UTAUT model approach in Jabodetabek. *International Journal of Artificial Intelligence Research, 6*(1). https://doi.org/10.29099/ijair.v6i1.282

Kinel, E., Korbel, K., Kozinoga, M., Czaprowski, D., Stępniak, Ł., & Kotwicki, T. (2021). The measurement of health-related quality of life of girls with mild to moderate idiopathic scoliosis: Comparison of ISYQOL versus SRS-22 questionnaire. *Journal of Clinical Medicine, 10*(21), 4806. https://doi.org/10.3390/jcm10214806

Kusmaryono, I., Wijayanti, D., & Maharani, H. R. (2022). Number of response options, reliability, validity, and potential bias in the use of the likert scale education and social science research: A literature review. *International Journal of Educational Methodology, 8*(4), 625–637. https://doi.org/10.12973/ijem.8.4.625

Lindner, J. R., & Lindner, N. (2024). Interpreting Likert type, summated, unidimensional, and attitudinal scales: I neither agree nor disagree, Likert or not. *Advancements in Agricultural Development, 5*(2), 152–163. https://doi.org/10.37433/aad.v5i2.351

Lord, F. M., & Novick, M. R. (2008). *Statistical theories of mental test scores*. IAP.

Mohamadi, Z. (2018). Comparative effect of online summative and formative assessment on EFL student writing ability. *Studies in Educational Evaluation, 59*(December), 29–40. https://doi.org/10.1016/j.stueduc.2018.02.003

Mustika, M., Maknun, J., & Feranie, S. (2019). Case study: Analysis of senior high school students' scientific creative, critical thinking and its correlation with their scientific reasoning skills on the sound concept. *Journal of Physics: Conference Series, 1157*(3), 1-5. https://doi.org/10.1088/1742-6596/1157/3/032057

Payan-Carreira, R., Sacau-Fontenla, A., Rebelo, H., Sebastião, L., & Pnevmatikos, D. (2022). Development and validation of a critical thinking assessment-scale short form. *Education Sciences, 12*(12) 938. https://doi.org/10.3390/educsci12120938

Putranta, H., & Supahar, S. (2019). Development of Physics-Tier Tests (PysTT) to measure students' conceptual understanding and creative thinking skills: A qualitative synthesis. *Journal for the Education of Gifted Young Scientists, 7*(3), 747–775. https://doi.org/10.17478/jegys.587203

Robitzsch, A. (2021). A comparison of linking methods for two groups for the two-parameter logistic item response model in the presence and absence of random differential item functioning. *Foundations, 1*(1), 116–144. https://doi.org/10.3390/foundations1010009

Robitzsch, A., & Lüdtke, O. (2020). A review of different scaling approaches under full invariance, partial invariance, and noninvariance for cross-sectional country comparisons in large-scale assessments. *Psychological Test and Assessment Modeling, 62*(2), 233–279.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1-36. https://doi.org/10.18637/jss.v048.i02

Şad, S. N. (2020). Does difficulty-based item order matter in multiple-choice exams? (Empirical evidence from university students). *Studies in Educational Evaluation, 64*(September), 100812. https://doi.org/10.1016/j.stueduc.2019.100812

Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika 38*(2), 203-219. https://doi.org/10.1007/BF02291114

Selçuk, E., & Demir, E. (2024). Comparison of item response theory ability and item parameters according to classical and Bayesian estimation methods. *International Journal of Assessment Tools in Education, 11*(2), 213–248. https://doi.org/10.21449/ijate.1290831

Shaw, A., Liu, O. L., Gu, L., Kardonova, E., Chirikov, I., Li, G., Hu, S., Yu, N., Ma, L., Guo, F., Su, Q., Shi, J., Shi, H., & Loyalka, P. (2020). Thinking critically about critical thinking: Validating the Russian HEIghten® critical thinking assessment. *Studies in Higher Education, 45*(9), 1933–1948. https://doi.org/10.1080/03075079.2019.1672640

Sidel, J. L., Bleibaum, R. N., & Tao, K. W. C. (2018). Quantitative descriptive analysis. In S. E. Kemp, J. Hort, & T. Hollowood (Eds.), *Descriptive analysis in sensory evaluation*. John Wiley & Sons Ltd. https://doi.org/10.1002/9781118991657.ch8

Sumintono, B., & Widhiarso, W. (2014). *Aplikasi model rasch untuk penelitian ilmu-ilmu sosial*. Trim Komunikata Publishing House.

Tobón, S., & Luna-nemecio, J. (2021). Complex thinking and sustainable social development: Validity and reliability of the complex-21 scale. *Sustainability, 13*(12), 1–19. https://doi.org/10.3390/su13126591

Tsikritsis, D., Legge, E. J., & Belsey, N. A. (2022). Practical considerations for quantitative and reproducible measurements with stimulated Raman scattering microscopy. *Analyst, 147*(21), 4642–4656. https://doi.org/10.1039/d2an00817c

Van Hauwaert, S. M., Schimpf, C. H., & Azevedo, F. (2020). The measurement of populist attitudes: Testing cross-national scales using item response theory. *Politics, 40*(1), 3–21. https://doi.org/10.1177/0263395719859306

Vollmer, F., & Alkire, S. (2022). Consolidating and improving the assets indicator in the global multidimensional poverty index. *World Development, 158*, 1–26, 105997. https://doi.org/10.1016/j.worlddev.2022.105997

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer International Publishing.

Widaman, K. F., & Revelle, W. (2023). Thinking thrice about sum scores, and then some more about measurement and analysis. *Behavior Research Methods, 55*(2), 788–806. https://doi.org/10.3758/s13428-022-01849-w

Zarate, D., Hobson, B. A., March, E., Griffiths, M. D., & Stavropoulos, V. (2023). Psychometric properties of the Bergen Social Media Addiction Scale: An analysis using item response theory. *Addictive Behaviors Reports, 17*(July), 100473. https://doi.org/10.1016/j.abrep.2022.100473

Zopluoglu, C. (2022). EstCRM: Calibrating Parameters for the Samejima's Continuous IRT Model. CRAN R Package. https://doi.org/10.32614/CRAN.package.EstCRM

Zou, G., Zou, L., & Qiu, S. fang. (2023). Parametric and nonparametric methods for confidence intervals and sample size planning for win probability in parallel-group randomized trials with likert item and likert scale data. *Pharmaceutical Statistics, 22*(3), 418–439. https://doi.org/10.1002/pst.2280