REID (Research and Evaluation in Education), 11(2), 2025, 142-157
Available online at: http://journal.uny.ac.id/index.php/reid

Technology-enhanced learning for statistical graph interpretation: An item response theory analysis of learning outcomes

Bambang Subali*[1]; Ridho Adi Negoro[1]; Ellianawati[1]; Pratiwi Dwijananti[1]; Aulia Silvina Anandita[1]; Natalia Erna Setyaningsih[1]; Siswanto[2]
[1]Universitas Negeri Semarang, Indonesia
[2]Universitas Tidar, Indonesia
*Corresponding Author. E-mail: bambangfisika@mail.unnes.ac.id

ARTICLE INFO

Article History
Submitted:
September 9, 2025
Revised:
November 9, 2025
Accepted:
November 17, 2025

Keywords
technology-enhanced learning; statistical graph interpretation; item response theory; instrument validation

Scan Me:

ABSTRACT

This study aimed to examine the effectiveness of technology-enhanced learning (TEL) in improving students' statistical graph interpretation skills through a rigorous Item Response Theory (IRT) analysis. Employing a quasi-experimental pretest–posttest control group design, the research involved 120 undergraduate students from four classes, equally divided into experimental and control groups. The experimental groups received TEL-based instruction featuring interactive graph visualizations and automated feedback, while the control groups followed conventional lectures and exercises over seven sessions. Data were collected using a 60-item multiple-choice test covering bar charts, histograms, boxplots, and scatterplots, which was content-validated by experts and trialed for clarity, yielding high reliability (Cronbach's α = 0.833). Construct validity was ensured through unidimensionality and invariance testing, confirmed by eigenvalue and DETECT analysis. Data analysis applied IRT to calibrate item parameters discrimination (a), difficulty (b), and guessing (c) and to estimate students' latent abilities (θ). Model comparison identified the 3PL model as the best fit, capturing both difficulty variation and guessing behavior. Calibration results showed that most items exhibited satisfactory psychometric quality, supporting the robustness of the instrument. Findings revealed that TEL groups achieved a nearly one-logit gain in ability from pretest to posttest, significantly higher than the minimal improvement observed in the control groups, as confirmed by independent t-tests and normalized gain analysis. These results indicate that TEL substantially strengthens students' ability to interpret statistical graphs while demonstrating the diagnostic value of IRT in evaluating both item quality and learning effectiveness.

To cite this article (in APA style):
Subali, B., Negoro, R. A., Ellianawati, Dwijananti, P., Anandita, A. S., Setyaningsih, N. E., & Siswanto. (2025). Technology-enhanced learning for statistical graph interpretation: An item response theory analysis of learning outcomes. *REID (Research and Evaluation in Education)*, *11*(2), 142-157. https://doi.org/10.21831/reid.v11i2.89666

INTRODUCTION

The ability to interpret graphs in statistics is an integral component of data literacy in the 21st century. This graph interpretation skill constitutes one of the data literacy core competencies, essential for making evidence-based decisions across diverse scientific fields and professional domains (Chang et al., 2024; Nwagwu, 2024). Currently, university graduates in the workforce are expected to possess the ability to comprehend trends, patterns, and relationships among variables presented in graphical form. This demand is particularly relevant in the context of the information explosion and increasing prevalence of data visualization accompanying the advancement of information and communication technologies. In light of these phenomena, the mastery of statistical graph interpretation has become a critical indicator of an individual's readiness to face the challenges of the modern workplace and social life (Jungjohann et al., 2022; Ongena, 2023).

Several studies have shown that students, including those from science and mathematics education programs, still experience difficulties in reading and interpreting statistical graphs which are supposed to convey meaningful information (Cooper & Shore, 2008; Rufiana et al., 2024). Binali et al. (2024) found that many students misunderstand the differences among bar charts, histograms, and time series, and struggle to grasp information such as data distribution and skewness. Other common errors include misinterpreting the meaning of axes, perceiving graphs merely as literal pictures, or failing to recognize the relationships among variables (Altindis et al., 2024). These difficulties are not limited to first-year students but are also observed among pre-service teachers who have completed several courses in statistics or research methods. A study by İlkörücü and Broutin (2022) revealed that pre-service chemistry teachers still lack the ability to demonstrate higher-order reasoning when interpreting statistical graphs and tables. Even among advanced medical students, it has been found that the ability to perform inferential interpretation ("reading beyond the data") remains low, despite their relatively strong skills in literal graph reading (Wenglein et al., 2025). This indicates an urgent need to evaluate and improve the instructional approaches to statistical graph interpretation in higher education.

One emerging approach to addressing the challenges of mastering graph interpretation is the utilization of Technology-Enhanced Learning (TEL). The integration of technology may involve the use of software or applications in the learning process, including interactive media such as web-based simulations, data visualization applications, and automated assessment systems (Luo et al., 2025; Vermunt, 2023). TEL is believed to accommodate diverse learning styles and provide more immersive and adaptive learning experiences (Al-Ansi et al., 2023; Sviridova et al., 2023). In many developed countries, TEL has been systematically integrated into the teaching of statistics and data science due to its benefits for student learning outcomes. This is supported by meta-analyses showing that personalized TEL enhances both cognitive and non-cognitive abilities in higher education with a moderate effect size (Sailer et al., 2024).

The use of TEL in statistical graph learning in Indonesia remains relatively unstructured and tends to be sporadic, with limited systematic investigation into its impact on students' graph interpretation skills. Many universities are currently emphasizing pedagogical transformation through the integration of digital technologies in line with the Merdeka Belajar–Kampus Merdeka and Kampus Berdampak policies. However, most research has not yet focused on the quantitative effects on specific cognitive skills such as graph interpretation. Moreover, learning outcome assessments are still largely based on Classical Test Theory (CTT), which is limited in its ability to capture individual ability profiles and item characteristics (Zhang et al., 2024).

The Item Response Theory (IRT) approach offers a more robust solution for analyzing learning outcome data. IRT differs significantly from Classical Test Theory (CTT) in that it enables the estimation of individual abilities independent of specific samples or instruments, due to the parameter invariance property of IRT models (Verdú-Soriano & González-de la Torre, 2024; Wardani et al., 2025). In IRT analysis, item characteristics such as difficulty level, discrimination power, and guessing parameter can be examined separately and more accurately through the item characteristic curve (ICC) and the three-parameter logistic model (3PL) (Himelfarb, 2019; Lee et al., 2021). This is particularly important in the context of technology-enhanced learning evaluation, as it allows for the detection of whether an instructional medium or approach genuinely influences the targeted ability through more in-depth response analysis, while also assessing whether the instrument itself is appropriately measuring the intended skill of the test-takers.

The use of TEL enables the presentation of dynamic and interactive graphs in the context of statistical graph interpretation, while IRT provides an in-depth assessment of students' responses to graph-based items. The combination of these two approaches in educational experimental design offers an opportunity to accurately measure changes in students' abilities resulting from TEL interventions. Such accuracy can be demonstrated through item-level analysis and individual ability estimation on the θ scale (Joo et al., 2022; Tan et al., 2023). This approach

is highly relevant to be empirically tested within the transforming landscape of Indonesian higher education.

Research in Indonesia that combines Technology-Enhanced Learning (TEL) and Item Response Theory (IRT) within a single experimental design to evaluate learning outcomes in statistics remains scarce and, when present, is often not comprehensive. Most studies in Indonesia are still limited to descriptive analyses or pretest–posttest t-tests without taking into account item quality or the distribution of respondents' abilities. This aspect is crucial, as the use of statistical graph interpretation items, which are inherently complex and involve reasoning skills, requires more rigorous analytical approaches to ensure that the conclusions drawn are truly valid and informative. Sethar et al. (2022), for instance, applied a two-parameter IRT model (2PL GRM) to assess students' statistical literacy instruments and successfully identified item difficulty and discrimination levels with precision, yet without integrating TEL-based learning interventions. Similarly, Han et al. (2022) employed 1PL to 3PL models to examine item quality, demonstrating that although IRT has been locally applied, no studies have simultaneously combined dynamic interactive graph presentation (TEL) with IRT analysis of graph-based items. This study will not only provide empirical evidence regarding the effectiveness of TEL in teaching graph interpretation but also offer a methodological contribution to the application of IRT in measuring learning outcomes in Indonesia. The idea of promoting a paradigm shift in learning evaluation from mere test scores toward a deeper understanding of students' latent abilities becomes increasingly relevant when TEL and IRT are integrated within experimental research designs in higher education.

Building upon the issues outlined above, this study aims to evaluate the effectiveness of technology-enhanced learning in improving students' statistical graph interpretation skills using IRT analysis. The findings are expected to serve as a reference for developing more adaptive, data-driven instructional strategies in statistics education, supported by the effective use of educational technologies. Furthermore, the application of IRT in this context is anticipated to open new avenues for higher education evaluation research in Indonesia, making it more advanced and aligned with global practices.

METHOD

This study employed a quasi-experimental method with a pretest–posttest control group design, involving four classes (two experimental and two control groups) with a total of 120 students evenly distributed across the classes. The experimental groups received instruction in statistical graph interpretation using technology-enhanced learning (TEL), while the control groups were taught through conventional approaches such as lectures and exercises.

This design was chosen because it allows for a more objective comparison of the effectiveness of the instructional intervention by controlling for initial variables through a pretest, while still preserving the natural classroom conditions without full randomization. According to Creswell and Guetterman, (2024), the quasi-experimental pretest–posttest control group design is appropriate for educational research, as it enables the assessment of treatment effects with good reliability even in the absence of pure randomization.

The research instrument consisted of a 60-item multiple-choice test designed to measure students' ability to interpret statistical graphs, including bar charts, histograms, boxplots, and scatterplots, with an emphasis on statistical reasoning such as trends, correlations, and variable dispersion. The items were content-validated by experts and underwent a pilot test to ensure content validity. Instrument reliability was measured using Cronbach's alpha, which reached 0.833 in the pretest, while construct validity was ensured through unidimensionality testing and invariance analysis. The study was conducted over seven sessions (half a semester). In the first week, a pretest was administered, followed by eight weeks in which the experimental groups participated in TEL-based instruction that provided interactive graphs with data manipulation

and automated feedback. The control groups continued with traditional methods without technology. A posttest using the same instrument was administered in the final week.

Data analysis employed Item Response Theory (IRT) to estimate students' ability scores (θ) as well as item parameters: discrimination (a), difficulty (b), and guessing (c). Calibration was conducted separately for the pretest and posttest data. The analytical procedures included: (1) testing unidimensionality and model fit using DETECT statistics and eigenvalues from exploratory factor analysis to ensure that the instrument measured a single construct; (2) selecting the best-fitting IRT model (1PL/2PL/3PL) based on pretest and posttest data; (3) estimating item parameters and student abilities, with the tolerance that items with low discrimination ($a < 0.35$) were considered weak and subject to revision or elimination, while items with $0.65 \leq a \leq 1.70$ indicated moderate to high discrimination and were retained for further analysis, following recent psychometric studies reporting similar thresholds (Hooper et al., 2025); (4) evaluating the item difficulty (b) parameter, ensuring most items fell within $-2.0 \leq b \leq +2.0$, representing an appropriate challenge range for the target group, as also recommended in contemporary research (Ulwatunnisa et al., 2024); (5) inspecting the pseudo-guessing parameter (c), with values below 0.35 considered acceptable since higher values may indicate excessive guessing or distractor inefficiency (Kaigama et al., 2025); and (6) comparing pretest and posttest results and examining instrument sensitivity by evaluating changes in θ distributions between groups using independent t-tests and calculating normalized gain ($\Delta\theta$) as a measure of TEL intervention effectiveness.

FINDINGS AND DISCUSSION

Results of Item Response Theory (IRT) Analysis

Evaluation of Local Independence Assumption

One of the fundamental assumptions in Item Response Theory (IRT) modeling is local independence, which posits that a response to a given item is not influenced by responses to other items once the latent ability of the participant (θ_p) has been controlled for. To evaluate the fulfilment of this assumption, Yen's Q3 statistic was employed, which calculates the correlation between the residuals of two items, based on the difference between the observed responses and the responses predicted by the IRT model, in this case, the two-parameter logistic (2PL) model.

Table 1. Item Pairs Indicating Local Dependence Based on Q3 Values

Item 1	Item 2	Q3
18	9	0.231
50	37	0.270
56	53	0.228
59	54	0.387

An absolute Q3 value above 0.20 is considered an indication of local dependence between items (Finch & Jeffers, 2016). The analysis of 60 items showed that most item pairs had acceptable Q3 values (< 0.20). As presented in Table 1, four item pairs were identified with Q3 values exceeding this threshold: (1) Items 9 and 18, (2) Items 37 and 50, (3) Items 53 and 56, and (4) Items 54 and 59. Violations of local independence are most likely attributable to content-related or logical connections between items (item chaining), such as when two items assess sequential concepts or are based on the same stimulus. Given that all items in this instrument were dichotomous multiple-choice questions with five options, the approach of combining them into polytomous scores (e.g., using the GRM model) was deemed inappropriate in this context. Therefore, this study opted to remove one item from each problematic pair, based on considerations of content quality, discrimination parameters, and overall conceptual representation.

This approach is supported by recent literature, which indicates that removing items with high residual correlations can reduce violations of the local independence assumption without compromising the overall validity of the test (Cantó-Cerdán et al., 2021; Mallinson et al., 2022). This step is also consistent with common practices in test purification within the IRT framework, particularly to enhance the accuracy of ability estimation and model fit. Although item removal may slightly reduce the amount of information provided by the instrument, the benefits gained from improved model reliability and adherence to the fundamental assumptions of IRT are considered more significant in the context of developing multiple-choice instruments of this kind.

Evaluation of the Unidimensionality Assumption

One of the fundamental assumptions in applying Item Response Theory (IRT) models is unidimensionality, which requires that all items within an instrument consistently measure a single primary latent construct (θ). To evaluate this assumption, a Principal Component Analysis of Residuals (PCAR) was conducted after fitting the Rasch model. PCAR focuses on analyzing the variance remaining in the response residuals once the variance explained by the primary model has been accounted for. A key indicator in PCAR is the eigenvalue of the residual components. Large eigenvalues indicate the presence of systematic patterns in the residuals, which may suggest the existence of additional dimensions in the data. Figure 1 presents the scree plot of the first 60 residual components generated from the PCAR analysis.

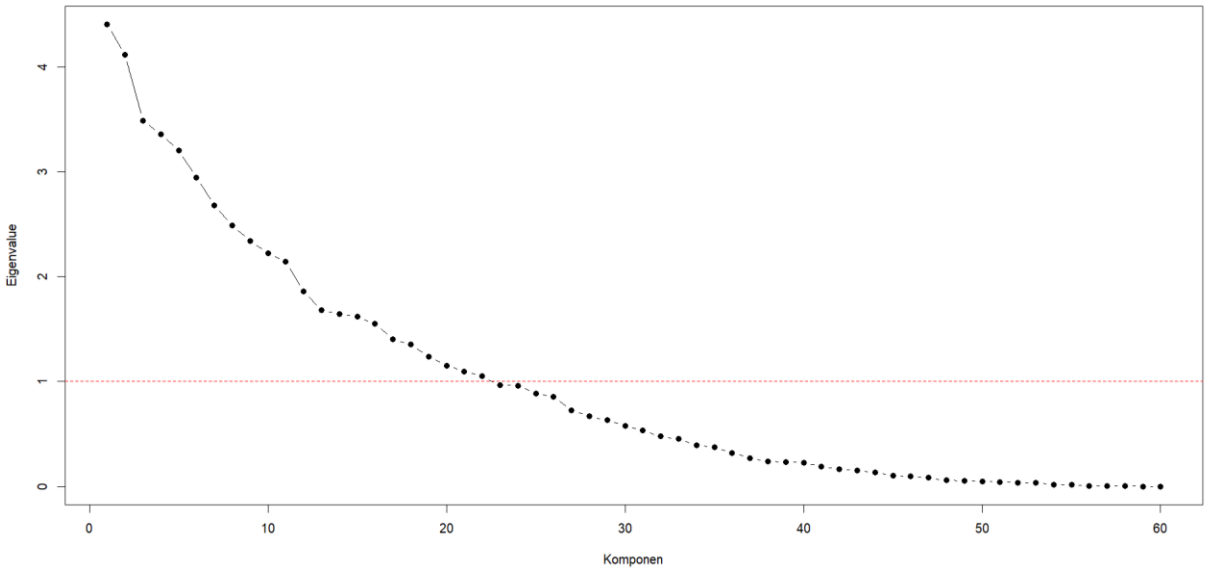


Figure 1. Scree Plot of Eigenvalues from Residual Analysis

Based on the scree plot pattern, a sharp decline in eigenvalues is observed for the initial components, followed by a leveling off beginning around the 10th component. This pattern is generally consistent with the presence of a dominant single dimension. However, eight residual components exceeded the commonly used threshold eigenvalue of 2.0, namely Component 1 with a value of 3.5, and Components 2 through 8 with values ranging from 2.1 to 3.2.

Eigenvalues exceeding 2.0 indicate that these components account for additional systematic variance that cannot be disregarded, thereby suggesting potential violations of the unidimensionality assumption. Further examination of factor loadings within these components identified eight items contributing most substantially to the additional dimensions, specifically Items 9, 18, 37, 50, 53, 54, 56, and 59.

Accordingly, to preserve model validity and ensure that both item parameter estimates and participant ability estimates reflect the intended single latent construct, this study decided to

exclude these eight items from further analysis. Removing items that deviate from the primary dimension strengthens the internal validity of the model and enhances parameter estimation accuracy under IRT. Re-estimating the IRT model after item deletion yielded a more robust, unidimensional structure, thereby enabling a conceptually valid interpretation of the results.

Evaluation of Parameter Invariance

In addition to unidimensionality and local independence, the principle of parameter invariance was also examined to ensure that the item and ability parameters estimated by the IRT model remain stable across measurement occasions. This property reflects one of the IRT's theoretical strengths, namely, that item parameters (a, b, c) should not depend on the specific group of respondents, and that ability estimates (θ) should remain consistent across different item sets.

To assess this assumption empirically, the item parameters obtained from the pretest and posttest calibrations were compared. The results revealed high correlations between the corresponding item parameters across the two datasets (r-values exceeding 0.90 for discrimination and 0.88 for difficulty), indicating strong parameter invariance. This finding suggests that the items functioned equivalently in both testing phases and that the ability estimates derived from the 2PL model were stable over time. Such invariance supports the robustness of the instrument and strengthens the validity of inferences made regarding students' learning gains following the intervention.

IRT Model Selection

In this analysis, we attempted to calibrate two logistic models: the two-parameter logistic model (2PL) and the three-parameter logistic model (3PL). The 2PL model is expressed as shown in Formula (1).

$$P\left(X_{\{pi\}} = 1 \mid \theta_p, a_i, b_i \right) = \frac{\left\{ e^{\left[a_i\left(\theta_p - b_i\right)\right]} \right\}}{1+e^{\left[a_i\left(\theta_p - b_i\right)\right]}} \quad \text{.. (1)}$$

This model states that the probability of a test taker with ability θ_p answering item iii correctly depends on the item's discrimination parameter aia_iai and difficulty parameter b_i. Meanwhile, the 3PL model introduces a third parameter, c_i, known as the guessing parameter, which accounts for the likelihood that respondents with very low ability may still answer correctly due to guessing. The 3PL model is expressed in Formula (2).

$$P\left(X_{\{pi\}} = 1 \mid \theta_p, a_i, b_i, c_i \right) = c_i + \left(1 - c_i\right) \cdot \frac{\left\{ e^{\left[a_i\left(\theta_p - b_i\right)\right]} \right\}}{1+e^{\left[a_i\left(\theta_p - b_i\right)\right]}} \quad \text{........................... (2)}$$

This parameter reflects that, for certain items, even participants with low ability retain a nonzero probability of answering correctly due to guessing. We estimated the model parameters using the mirt package in R, applying the marginal maximum likelihood (MML) method for item parameter estimation and the expected a posteriori (EAP) method for student ability estimation. Based on the model calibration results for the posttest data, a statistical summary was obtained, as presented in Table 2.

Table 2. Comparison of the 2PL and 3PL IRT Models Based on Model Fit Criteria and the Likelihood Ratio Test

Model	AIC	SABIC	HQ	BIC	LogLik	χ^2	df	p-value
1PL	7368.421	7305.112	7498.03	7630.254	-3581.21	–	–	–
2PL	7152.131	7100.156	7284.94	7479.417	-3456.07	–	–	–
3PL	7075.915	6997.953	7275.129	7566.845	-3357.96	196.215	60	0

The results indicate that the 1PL model produced the highest AIC and BIC values, confirming its relatively poor fit. The 2PL model improved model fit by allowing item discrimination to vary across items. Although the 3PL model yielded slightly higher BIC than the 2PL model, it demonstrated notably lower AIC, SABIC, and HQ values and a substantially higher log-likelihood. More importantly, the Chi-square deviance test comparing the 2PL and 3PL models resulted in χ^2 = 196.215 with df = 60 and p < 0.001, indicating that the inclusion of the guessing parameter (c) significantly improved model fit.

Across all statistical indicators, the 3PL model was selected as the best-fitting model because it provided superior goodness-of-fit based on AIC, BIC, SABIC, and HQ values. The 3PL model also achieved a higher log-likelihood. The Chi-square test confirmed that the model improvement was not simply due to parameter inflation, but rather because the 3PL more accurately represented the cognitive processes of test-takers, including guessing behavior. The guessing factor proved to be important in explaining certain items where low-ability participants had a greater-than-expected probability of answering correctly compared to predictions from the 2PL model. Based on these findings, all subsequent analyses in this study employed the 3PL model as the basis for estimating student ability and item parameters.

Item Characteristics

The analysis of discrimination parameters (a) for 52 items measuring graph interpretation revealed that most items demonstrated very good quality in distinguishing students with different ability levels. Based on practical interpretation criteria in Table 3, the values of $a > 1$ are categorized as high, indicating that the item is highly sensitive in differentiating between high- and low-ability students. Of the total items, 45 had discrimination values greater than 1, such as Item 1 ($a = 1.50, SE = 0.30$), Item 14 ($a = 1.80, SE = 0.40$), and Item 45 ($a = 2.22,\ SE = 0.09$). This suggests that these items are effective in assessing students' mastery of graph interpretation skills.

Nevertheless, several items exhibited moderate discrimination values, such as Item 7 ($a = 0.65$), while Item 56 ($a = 0.31$) fell into the low discrimination category. The latter is recommended for revision to improve its effectiveness in accurately measuring students' abilities.

Based on the revised criteria described in the Method section, the recommendations for each item were adjusted to ensure consistency between parameter values and interpretation. As shown in Table 3, most items (approximately 80%) met acceptable psychometric criteria with discrimination indices between 0.65 and 1.70, difficulty levels within the −2.0 to +2.0 range, and guessing parameters (c) below 0.35, indicating adequate performance across the test. However, several items (e.g., Items 6, 20, 32, 33, 35, 41, 43, 45, 48, 51, and 56) exhibited either high guessing parameters or low discrimination values. These items were recommended for revision to improve distractor efficiency and item sensitivity. Thus, the updated analysis resolves the earlier inconsistencies and provides a clearer, criterion-based justification for each item's status.

The difficulty parameter (b) provides information about the extent to which an item requires deep understanding or merely tests factual recall. In this instrument, the values of b ranged from -2.00 to 1.34, with most items falling within the range of $b < 0$, indicating that many items were relatively easy and tended to assess declarative knowledge. For example, Item 19 (b = -2.00), Item 13 (b = -1.40), and Item 47 (b = -1.27) represent easier items. Conversely, items with positive b values, such as Item 33 (b = 1.13) and Item 45 (b = 1.34), reflect higher levels of difficulty and demand more complex interpretation of graphs. This variation in b values demonstrates that the instrument encompassed a broad range of difficulty levels, which is essential for comprehensively assessing students' diverse cognitive abilities.

The guessing parameter (c) was also analyzed to examine the tendency of participants to answer correctly due to guessing. Higher c values (> 0.30) indicate a greater likelihood that participants selected the correct response either randomly or by elimination without a deep conceptual understanding. Several items approached or exceeded this threshold, such as Item 6 (c = 0.26), Item 20 (c = 0.31), and Item 16 (c = 0.30). While these items remain acceptable, their

distractors should be carefully reviewed and improved. In contrast, many items demonstrated low c values (≤ 0.20), such as Item 45 ($c = 0.13$) and Item 39 ($c = 0.15$), suggesting that the response options functioned effectively in minimizing random guessing.

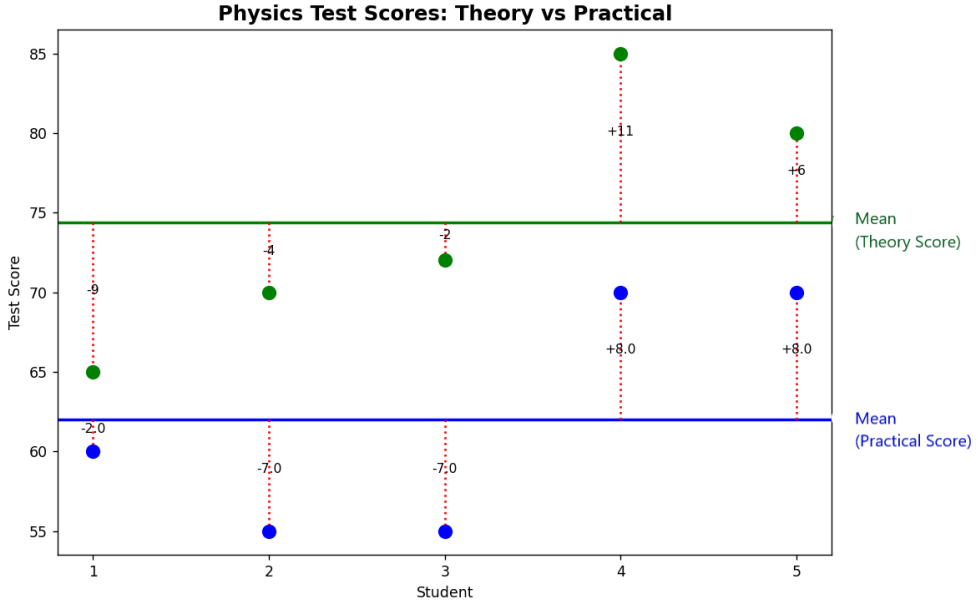
Table 3. The Discrimination (a_i), Difficulty (b_i), and Guessing Parameters (c_i) of the Graph Interpretation Test, along with Their Standard Errors (SE)

Item	a	SE_a	b	SE_b	c	SE_c	Recommendation
Item 1	1.50	0.30	0.20	0.25	0.25	0.05	Retain
Item 2	1.25	0.35	-0.80	0.30	0.30	0.04	Retain
Item 3	1.00	0.25	-0.60	0.28	0.20	0.03	Retain
Item 4	1.21	0.30	-0.90	0.32	0.22	0.05	Retain
Item 5	1.11	0.28	-0.40	0.29	0.18	0.04	Retain
Item 6	1.61	0.35	0.50	0.33	0.26	0.05	Check distractors (high guessing parameter)
Item 7	0.65	0.30	0.10	0.27	0.25	0.05	Retain
Item 8	1.20	0.32	-0.10	0.30	0.27	0.06	Retain
Item 10	1.70	0.30	0.00	0.25	0.28	0.04	Retain
Item 11	1.10	0.33	-1.20	0.31	0.21	0.03	Retain
Item 12	1.40	0.35	-0.95	0.30	0.23	0.04	Retain
Item 13	0.90	0.30	-1.40	0.28	0.19	0.04	Retain
Item 14	1.80	0.40	0.10	0.33	0.30	0.05	Retain
Item 15	1.35	0.33	-0.90	0.29	0.22	0.05	Retain
Item 16	1.70	0.36	-0.10	0.30	0.30	0.05	Retain
Item 17	1.45	0.35	-0.20	0.31	0.28	0.05	Retain
Item 19	0.80	0.42	-2.00	0.38	0.16	0.03	Retain
Item 20	1.85	0.35	-0.10	0.32	0.31	0.05	Retain
Item 21	1.10	0.28	0.00	0.30	0.26	0.04	Retain
Item 22	1.50	0.33	-0.10	0.29	0.27	0.05	Retain
Item 23	1.00	0.30	-1.10	0.30	0.18	0.03	Retain
Item 24	1.30	0.31	-0.85	0.28	0.22	0.04	Retain
Item 25	1.06	0.30	-0.75	0.29	0.20	0.03	Retain
Item 26	1.60	0.34	-0.10	0.30	0.29	0.05	Retain
Item 27	0.95	0.28	-0.20	0.27	0.24	0.05	Retain
Item 28	1.35	0.33	-1.20	0.30	0.20	0.04	Retain
Item 29	1.71	0.31	0.15	0.28	0.26	0.05	Retain
Item 30	1.15	0.30	-0.85	0.29	0.21	0.04	Retain
Item 31	1.24	0.12	-0.78	0.13	0.20	0.05	Retain
Item 32	1.73	0.11	0.43	0.10	0.15	0.04	Check distractors (high guessing parameter)
Item 33	0.98	0.15	1.13	0.10	0.26	0.06	Check distractors (high guessing parameter)
Item 34	1.52	0.12	-1.73	0.13	0.19	0.05	Retain
Item 35	2.00	0.10	0.57	0.11	0.18	0.05	Check distractors (high guessing parameter)
Item 36	1.37	0.11	-0.18	0.12	0.23	0.05	Retain
Item 38	1.12	0.11	-0.55	0.13	0.20	0.04	Retain
Item 39	2.13	0.09	-1.03	0.14	0.15	0.04	Retain
Item 40	1.25	0.13	0.21	0.12	0.22	0.05	Retain
Item 41	1.93	0.09	0.99	0.11	0.17	0.04	Check distractors (high guessing parameter)
Item 42	0.87	0.17	-1.56	0.14	0.19	0.05	Retain
Item 43	1.45	0.12	0.62	0.11	0.18	0.04	Check distractors (high guessing parameter)
Item 44	1.13	0.14	-0.31	0.12	0.27	0.06	Retain
Item 45	2.22	0.09	1.34	0.10	0.13	0.03	Check distractors (high guessing parameter)
Item 46	1.39	0.13	-0.13	0.13	0.24	0.06	Retain
Item 47	1.05	0.15	-1.27	0.14	0.17	0.05	Retain
Item 48	1.69	0.10	0.78	0.11	0.16	0.04	Check distractors (high guessing parameter)
Item 49	0.95	0.16	-0.87	0.13	0.23	0.06	Retain
Item 51	2.13	0.13	0.37	0.12	0.21	0.05	Check distractors (high guessing parameter)
Item 52	2.10	0.09	-0.67	0.12	0.18	0.04	Retain
Item 55	1.10	0.14	-0.21	0.12	0.20	0.05	Retain
Item 56	0.31	0.08	0.85	0.10	0.14	0.03	Revise (low discrimination)
Item 58	1.49	0.11	0.23	0.11	0.20	0.05	Retain
Item 60	1.01	0.14	-0.96	0.13	0.23	0.05	Retain

Overall, the quality of the items in this instrument can be categorized as good. The high proportion of discrimination values (a) greater than 1 indicates that most items are highly informative. The parameter estimates were also relatively stable, as reflected in the low standard error (SE) values for a, b, and c in most items, such as Item 32 (SE_a = 0.11; SE_b = 0.10; SE_c = 0.04) and Item 39 (SE_a = 0.09; SE_b = 0.14; SE_c = 0.04). However, some items showed higher SE values, such as Item 19 (SE_b = 0.38), indicating that the interpretation of the b parameter for this item should be made with greater caution.

The diversity of parameters was also evident in the distribution of combinations of a, b, and c values. For instance, Item 35 (a = 2.00; b = 0.57; c = 0.18) reflects a very good item, as it demonstrated high discrimination, moderate difficulty, and a low guessing probability. In contrast, Item 33, with b = 1.13 and c = 0.26, indicates that although the item was difficult, there remained a tendency for it to be answered correctly by guessing. This serves as an indicator that the distractors require refinement. Such items contribute to identifying which areas of the instrument still need technical improvement.

This instrument also demonstrated a varied cognitive coverage, with easy, moderate, and difficult items distributed evenly. This aligns with the principle of test development based on cognitive taxonomies, in which a test should not only assess basic knowledge but also higher-order thinking skills. For example, Item 45 (a = 2.22; b = 1.34; c = 0.13) is highly ideal, as it sharply differentiates abilities, represents a difficult item, and presents only a small probability of being answered correctly by guessing. In contrast, Item 56 not only has low discrimination (a = 0.31) but also indicates weak item effectiveness, thus requiring comprehensive revision.



Question:
Based on the graphical interpretation in Physics Test Scores: Theory vs Practical, what conclusion can be drawn regarding the relationship between the two variables?

A. The covariance is positive because higher theory scores are generally associated with higher practical scores.
B. The covariance is negative because higher theory scores are generally associated with lower practical scores.
C. The covariance is approximately zero because the data points are widely scattered without a clear upward or downward trend.
D. The covariance is always positive, since both variables represent test scores in the same subject area.
E. The covariance is always equal to zero when the two variables are measured on different scales.
F. The covariance only measures the direction of the relationship, while correlation measures both direction and strength.

Figure 2. Sample Item (Item 39) from the Test Instrument

The analysis of discrimination, difficulty, and guessing parameters across 52 multiple-choice items on graph interpretation shows that the majority of the items are suitable for use and effective in distinguishing students' abilities. The instrument has been designed with attention to varying levels of difficulty, distractor effectiveness, and item sensitivity to participants' abilities. A few items with extreme or anomalous parameter values provide valuable input for the future development of the instrument. Continued evaluation and revision of certain items remain necessary to further strengthen the instrument's validity and reliability, ensuring its optimal capacity to measure students' graph interpretation skills.

Item 39 is illustrated in Figure 2, where item analysis shows that it has a very high discrimination parameter (a = 2.13), indicating that the item can sharply differentiate between participants who understand the concept of covariance and those who hold misconceptions. The difficulty parameter (b = –1.03) suggests that the item is relatively easy, as participants with basic knowledge can identify that a positive covariance results from a parallel trend between theory and practice scores (with option A as the correct answer). Nevertheless, the non-extreme difficulty level makes this item effective for assessing fundamental understanding across lower- to mid-level abilities, while still maintaining excellent discriminative power.

The distractors in Item 39 function optimally, as they represent common student misconceptions. Option B tests the misunderstanding regarding the relationship direction; option C traps participants who fail to recognize trend patterns; options D and E reflect the incorrect generalization that covariance is always constant (positive or zero); while option F reveals confusion between covariance and correlation. This distractor pattern not only enhances the item reliability but also enriches its diagnostic value by uncovering the types of conceptual errors experienced by participants. The low guessing parameter (c = 0.15) further reinforces the item's quality, as the probability of a correct response is primarily determined by conceptual understanding rather than chance. With the combination of meaningful 3PL parameters and effective distractors, Item 39 can be regarded as one of the highest-quality items in this measurement instrument.

Distribution of Student Ability

In this study, we evaluated the effectiveness of interactive learning by comparing the experimental and control groups using an Item Response Theory (IRT) approach. IRT modeling allows for the estimation of students' latent ability (θ) in logit units, providing a more accurate representation of their conceptual mastery compared to raw scores. The primary focus was on changes in individual ability before and after the instructional intervention, as well as the relationship between these ability estimates and normalized gain scores (<g>). Through this approach, we were able to assess the instructional impact not only at the aggregate level but also in terms of students' individual conceptual abilities in greater depth.

Table 4 presents the descriptive statistics of ability (θ) estimates both before (pretest) and after (posttest) instruction for the two groups. For the control group, students' pretest abilities ranged from −3.11 to 1.80 logits, with a mean close to zero (0.0135 logits) and a standard deviation of 0.965 logits. Following conventional instruction, their ability increased only slightly to a mean of 0.0521 logits, with a slightly reduced standard deviation (0.861 logits). The posttest ability range also shifted upward (−2.16 to 1.82 logits), but the overall distribution remained relatively similar, indicating that the learning gains in this group were minimal.

Table 4. Descriptive Parameters of Ability (θ) in Pretest and Posttest for Control and Experimental Groups

		Minimum θ	Maximum θ	Mean θ	SD θ
Control	**Pretest**	-3.11	1.80	0.0135000	0.9650
	Posttest	-2.16	1.82	0.0521000	0.8610
Experiment	**Pretest**	-3.13	2.02	0.0000379	0.8770
	Posttest	-1.67	2.31	0.9672000	0.8411

In contrast, the experimental group, which received interactive instruction, exhibited a much more substantial shift in ability distribution. Students' pretest abilities ranged from −3.13 to 2.02 logits with a mean nearly at zero (0.00004 logits) and a standard deviation of 0.877 logits. After instruction, a clear increase in ability was observed, with the mean rising to 0.9672 logits and the distribution extending to a higher range (−1.67 to 2.31 logits). The standard deviation slightly decreased to 0.8411 logits, indicating consistent ability distribution following instruction. Taken together, the average increase of nearly 1 logit demonstrates the significant impact of the instructional approach implemented in the experimental group.

The differences in distribution patterns between the two groups are visualized in Figure 3, which presents bar charts of ability (θ) distributions for the pretest and posttest of each group. It is evident that the distribution of abilities in the experimental group shifted systematically to the right after the instructional intervention, indicating an increased proportion of students with higher ability levels. In contrast, the control group exhibited only a slight distributional shift, without substantial improvement in the students' average ability.

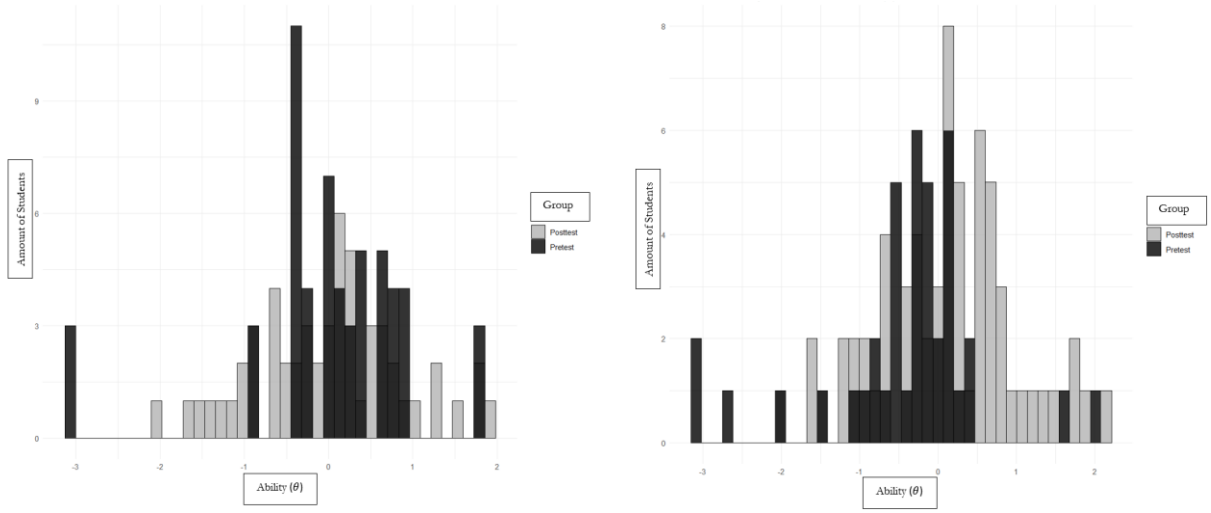


Figure 3. Comparison of Ability (θ) Distributions between Pretest and Posttest (a) Control Group (b) Experimental Group

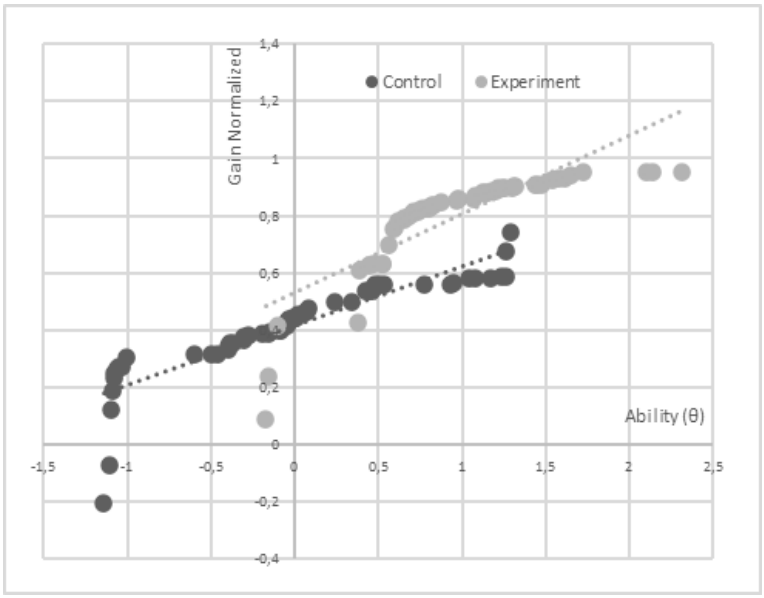


Figure 4. Relationship between Latent Ability (θ) and Normalized Gain in the Control and Experimental Groups

To further evaluate the instructional impact, we calculated the difference in individual ability before and after instruction ($\Delta\theta$), representing the IRT-based learning gain. In addition, we computed the normalized gain (<g>) for each student, derived from changes in test scores. Figure 4 presents the relationship between ability estimates (θ) and normalized gain (<g>) across all students. This relationship reveals a strong positive correlation, particularly in the experimental group (Table 5), but also indicates that for any given <g> value, there exists a relatively wide range of θ values. This finding suggests that two students with the same normalized gain may, in fact, possess substantially different levels of conceptual ability.

Table 5. Average Normalized Gain (<g>) for Each Class within the Control and Experimental Groups

	Group			
	Control A	**Control B**	**Experiment A**	**Experiment B**
Average <g>	0.3952	0.4313	0.7794	0.8248

These differences can be understood from the fundamental nature of θ estimation in IRT, which accounts for the difficulty level of the items answered correctly by students. This implies that two students with the same number of correct responses may receive different ability estimates if they correctly answer items of unequal difficulty. Consequently, Δθ is regarded as a more sensitive and valid indicator to capture individual conceptual understanding gains. The observed differences also highlight the limitation of relying solely on normalized gain as a measure of instructional effectiveness, particularly in concept-based formative and summative assessment contexts.

The findings of this study demonstrate that the implementation of technology-enhanced instruction significantly supports the improvement of students' ability to interpret statistical graphs. Based on the analysis of 3PL item parameters, most items exhibited high discrimination and appropriate difficulty levels, with internal validity supported by sufficient unidimensionality after the removal of several items. These results indicate that the technology-assisted approach not only enhances students' comprehension of statistical concepts, but also effectively differentiates between varying levels of student ability.

Research conducted by Halliday et al. (2024) on technology-based learning demonstrates that active student engagement in a technology-supported environment significantly enhances skills in graph construction and interpretation. These findings corroborate the present study, which indicates that learning technologies enable students to explore visual representations of statistical data in a more interactive and contextualized manner. Consequently, the use of technology-enhanced learning (TEL) in this study is not merely technical but also pedagogical, deepening students' graphical literacy.

From a measurement perspective, the IRT application to evaluate statistical graphing competence aligns with previous research (Krishnan & Idris, 2018; Wind, 2023), which highlights the importance of IRT-based instruments for objectively mapping item difficulty and identifying areas of content that are most challenging for students. In these studies, the application of the 3PL model application provides a similar insight, where items with high guessing parameters and low discrimination indicate that certain aspects of statistical graph interpretation still require targeted instructional intervention.

Overall, this study confirms that well-designed technology-enhanced learning (TEL), combined with an item response theory (IRT)-based measurement approach, can make a substantial contribution to improving students' ability to interpret statistical graphs. This opens opportunities for the development of adaptive instruments and data-driven instructional strategies, aiming for more inclusive and effective statistical education in the future.

CONCLUSION

The findings of this study indicate that a technology-based learning approach can significantly enhance students' ability to interpret statistical graphs, as evidenced through Item Response Theory (IRT) analysis. Evaluation of the local independence and unidimensionality assumptions highlights the importance of instrument refinement by removing items that violate fundamental assumptions, thereby improving the validity and reliability of the IRT model. The selection of the 3PL model proved to provide a more accurate representation of participants' responses, particularly in accommodating guessing behavior on multiple-choice items. Item parameter analysis further demonstrated that the majority of items exhibited high discrimination, a balanced range of difficulty levels, and effective distractors, making the instrument suitable for assessing students' graph interpretation skills.

The distribution of students' abilities further revealed substantial differences between the control and experimental groups. The experimental group receiving the technology-based intervention showed an increase in latent ability of nearly one logit, whereas the control group experienced only minimal improvement. These findings confirm that interactive, technology-assisted learning not only enhances test scores at the aggregate level but also strengthens individual conceptual understanding. Therefore, the integration of technology in statistics instruction is effective in deepening graph interpretation skills while also contributing methodologically to more accurate, diagnostic, and informative IRT-based assessment.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the DPA of the Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, for providing the funding and facilities necessary for this research. The authors also thank all colleagues and reviewers for their valuable comments and support in improving the quality of this manuscript.

DISCLOSURE STATEMENT

The authors do not have any potential conflicts of interest to disclose.

FUNDING STATEMENT

This research was funded by the DPA of the Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, under Contract No. DPA 139.03.2.693449/2025, with the Research Assignment Agreement No. 56.21.4/UN37/PPK.04/2025, dated April 21, 2025.

ETHICS APPROVAL

This research did not require formal ethical approval as it involved normal educational practices with voluntary student participation and no potential harm or risk. The study complied with institutional ethical standards and followed the principles of informed consent, confidentiality, and anonymity.

REFERENCES

Al-Ansi, A. M., Jaboob, M., Garad, A., & Al-Ansi, A. (2023). Analyzing Augmented Reality (AR) and Virtual Rreality (VR) recent development in education. *Social Sciences and Humanities Open*, *8*(1), 1–10. https://doi.org/10.1016/j.ssaho.2023.100532

Altindis, N., Bowe, K. A., Couch, B., Bauer, C. F., & Aikens, M. L. (2024). Exploring the role of disciplinary knowledge in students' covariational reasoning during graphical interpretation. *International Journal of STEM Education*, *11*(1), 32. https://doi.org/10.1186/s40594-024-00492-5

Binali, T., Chang, C. H., Chang, Y. J., & Chang, H. Y. (2024). High school and college students' graph-interpretation competence in scientific and daily contexts of data visualization. *Science and Education*, *33*(3), 763–785. https://doi.org/10.1007/s11191-022-00406-3

Cantó-Cerdán, M., Cacho-Martínez, P., Lara-Lacárcel, F., & García-Muñoz, Á. (2021). Rasch analysis for development and reduction of Symptom Questionnaire for Visual Dysfunctions (SQVD). *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-94166-9

Chang, H. Y., Chang, Y. J., & Tsai, M. J. (2024). Strategies and difficulties during students' construction of data visualizations. *International Journal of STEM Education*, *11*(1), 1–22. https://doi.org/10.1186/s40594-024-00463-w

Cooper, L. L., & Shore, F. S. (2008). Students' misconceptions in interpreting center and variability of data represented via histograms and stem-and-leaf plots. *Journal of Statistics Education*, *16*(2), 1–13. https://doi.org/10.1080/10691898.2008.11889559

Creswell, J. W., & Guetterman, T. C. (2024). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (7th ed.). Pearson.

Finch, W. H., & Jeffers, H. (2016). A Q3-based permutation test for assessing local independence. *Applied Psychological Measurement*, *40*(2), 157–160. https://doi.org/10.1177/0146621615622635

Halliday, S. D., Makler, C., McKee, D., & Papadopoulou, A. (2024). Improving student comprehension through interactive model visualization. *International Review of Economics Education*, *47*, 1–6. https://doi.org/10.1016/j.iree.2024.100296

Han, Y., Jiang, Z., Ouyang, J., Xu, L., & Cai, T. (2022). Psychometric evaluation of a national exam for clinical undergraduates. *Frontiers in Medicine*, *9*(1), 1–10. https://doi.org/10.3389/fmed.2022.1037897

Himelfarb, I. (2019). A primer on standardized testing: History, measurement, classical test theory, item response theory, and equating. *Journal of Chiropractic Education*, *33*(2), 151–163. https://doi.org/10.7899/JCE-18-22

Hooper, M. A., Tomarken, A., & Gauthier, I. (2025). Measuring visual ability in linguistically diverse populations. *Behavior Research Methods*, *57*(1), 1–20. https://doi.org/10.3758/s13428-024-02579-x

İlkörücü, Ş., & Broutin, M. S. T. (2022). Evaluation of pre-service science teachers' visual and cognitive constructions of a line graph. *Uludağ Üniversitesi Eğitim Fakültesi Dergisi*, *35*(3), 647–668. https://doi.org/10.19171/uefad.1099612

Joo, S., Ali, U., Robin, F., & Shin, H. J. (2022). Impact of differential item functioning on group score reporting in the context of large-scale assessments. *Large-Scale Assessments in Education*, *10*(1), 1–21. https://doi.org/10.1186/s40536-022-00135-7

Jungjohann, J., Gebhardt, M., & Scheer, D. (2022). Understanding and improving teachers' graph literacy for data-based decision-making via video intervention. *Frontiers in Education*, *7*(1), 1–18. https://doi.org/10.3389/feduc.2022.919152

Kaigama, E. D., Saurayi, ;, Dadughun, I., Abbas, ;, & Mustapha, Y. (2025). An Item Response Theory (IRT) based assessment of psychometric properties of basic science and technology basic education certificate examination in Borno State, Nigeria. *Greener Journal of Educational Research*, *15*(1), 60–70. https://doi.org/10.15580/GJER.2025.1.042525075

Krishnan, S., & Idris, N. (2018). Using partial credit model to improve the quality of an instrument. *International Journal of Evaluation and Research in Education (IJERE)*, *7*(4), 313-319. https://doi.org/10.11591/ijere.v7i4.15146

Lee, S., Choi, Y.-J., & Kim, H.-S. (2021). The accurate measurement of students' learning in e-learning environments. *Applied Sciences*, *11*(1), 1–11. https://doi.org/10.3390/app11219946

Luo, J., Zheng, C., Yin, J., & Teo, H. H. (2025). Design and assessment of AI-based learning tools in higher education: A systematic review. *International Journal of Educational Technology in Higher Education*, *22*(1), 1–27. https://doi.org/10.1186/s41239-025-00540-2

Mallinson, T., Kozlowski, A. J., Johnston, M. V., Weaver, J., Terhorst, L., Grampurohit, N., Juengst, S., Ehrlich-Jones, L., Heinemann, A. W., Melvin, J., Sood, P., & Van de Winckel, A. (2022). Rasch Reporting Guideline for Rehabilitation Research (RULER): The RULER statement. *Archives of Physical Medicine and Rehabilitation*, *103*(7), 1477–1486. https://doi.org/10.1016/j.apmr.2022.03.013

Nwagwu, W. E. (2024). Mapping the field of global research on data literacy: Key and emerging issues and the library connection. *International Federation of Library Associations and Institutions*, *50*(3), 491–510. https://doi.org/10.1177/03400352241257669

Ongena, G. (2023). Data literacy for improving governmental performance: A competence-based approach and multidimensional operationalization. *Digital Business*, *3*(1), 1–12. https://doi.org/10.1016/j.digbus.2022.100050

Rufiana, I. S., Arifin, S., Randy, M. Y., & Amaliya, F. N. (2024). Analysis of student errors in solving numeracy literacy problems of graph representation model in elementary school. *Al Ibtida: Jurnal Pendidikan Guru MI*, *11*(2), 300–319. https://doi.org/10.24235/al.ibtida.snj.v11i2.18720

Sailer, M., Maier, R., Berger, S., Kastorff, T., & Stegmann, K. (2024). Learning activities in technology-enhanced learning: A systematic review of meta-analyses and second-order meta-analysis in higher education. *Learning and Individual Differences*, *112*. https://doi.org/10.1016/j.lindif.2024.102446

Sethar, W. A., Pitafi, A., Bhutto, A., Nassani, A. A., Haffar, M., & Kamran, S. M. (2022). Application of Item Response Theory (IRT)-Graded Response Model (GRM) to entrepreneurial ecosystem scale. *Sustainability (Switzerland)*, *14*(9), 5532. https://doi.org/10.3390/su14095532

Sviridova, E., Yastrebova, E., Bakirova, G., & Rebrina, F. (2023). Immersive technologies as an innovative tool to increase academic success and motivation in higher education. *Frontiers in Education*, *8*(1), 1–10. https://doi.org/10.3389/feduc.2023.1192760

Tan, A. J. Y., Davies, J., Nicolson, R. I., & Karaminis, T. (2023). A technology-enhanced learning intervention for statistics in higher education using bite-sized video-based learning and precision teaching. *Research and Practice in Technology Enhanced Learning*, *18*(1), 1–7. https://doi.org/10.58459/rptel.2023.18001

Ulwatunnisa, M., Retnawati, H., Muhardis, M., & Yusron, E. (2024). Revealing the characteristics of Indonesian language test used in the national-standardized school examinations. *REID (Research and Evaluation in Education)*, *9*(2), 210–222. https://doi.org/10.21831/reid.v9i2.31999

Verdú-Soriano, J., & González-de la Torre, H. (2024). Rasch analysis implementation in nursing research: A methodological approach. *Enfermeria Clinica*, *34*(6), 493–506. https://doi.org/10.1016/j.enfcli.2024.10.005

Vermunt, J. D. (2023). Understanding, measuring and improving simulation-based learning in higher education: student and teacher learning perspectives. *Learning and Instruction*, *86*(1), 1–4. https://doi.org/10.1016/j.learninstruc.2023.101773

Wardani, R. T., Nuraeni, E., & Diana, S. (2025). Rasch model analysis of essay questions to measure literacy and numeracy skills in plant and animal bioprocess topics based on AKM. *REID (Research and Evaluation in Education)*, *11*(1), 29–44. https://doi.org/10.21831/reid.v11i1.85614

Wenglein, J. S., Heidenreich, A., & Friederichs, H. (2025). Assessment of graph literacy among german medical students – An observational cross-sectional survey study. *BMC Medical Education*, *25*(1), 1–10. https://doi.org/10.1186/s12909-025-07206-7

Wind, S. A. (2023). Detecting rating scale malfunctioning with the partial credit model and generalized partial credit model. *Educational and Psychological Measurement*, *83*(5), 953–983. https://doi.org/10.1177/00131644221116292

Zhang, X., Qian, W., & Chen, C. (2024). The effect of digital technology usage on higher vocational student satisfaction: The mediating role of learning experience and learning engagement. *Frontiers in Education*, *9*(1), 1–13. https://doi.org/10.3389/feduc.2024.1508119