Jurnal Elemen, 11(1), 153-170, January 2025
https:/doi.org/10.29408/jel.v11i1.27634

e- ISSN: 2442 - 4226

Check for updates

Integration of item response theory in the development of PhET-based graphing lines worksheets for optimizing student algebra competence

Giyanti, Indri Lestari, Rina Oktaviyanthi *

Department of Mathematics Education, Universitas Serang Raya, Banten, Indonesia

* Correspondence: rinaokta@unsera.ac.id

Abstract

This study develops and evaluates a graphing line worksheet based on PhET Interactive Simulation integrated with Item Response Theory (IRT) methods to enhance student algebra competence. Involving 120 students, the worksheet comprises 12 items measuring four key indicators: understanding the geometric significance of line slopes, constructing line equations, graphing from line equations, and predicting the effects of variable changes. The 2-Parameter Logistic (2PL) model of IRT was employed to analyze item difficulty and student ability in logit form. The results indicate that the worksheet is effective in improving student algebra competence, with Items 1 and 11 demonstrating a good balance between difficulty and discrimination. Item 2 requires further review because of its high difficulty, whereas Item 12 is considered too easy. Heatmap analysis and Item Characteristic Curves (ICC) revealed variations in student response patterns, confirming the test's ability to evaluate diverse levels of student ability. The integration of interactive simulation and IRT has proven to be an effective strategy in instructional design, supporting adaptive and personalized learning.

Keywords: algebra competence; graphing lines; item response theory; parameter logistic; phet interactive simulation

How to cite: Giyanti, Lestari, I., & Oktaviyanthi, R. (2025). Integration of item response theory in the development of PhET-based graphing lines worksheets for optimizing student algebra competence. *Jurnal Elemen*, *11*(1), 153-170. https://doi.org/10.29408/jel.v11i1.27634

Received: 30 September 2024 | Revised: 12 November 2024
Accepted: 8 December 2024 | Published: 1 February 2025

Introduction

Algebra competence is a fundamental skill that students must acquire early on due to its crucial role in developing analytical thinking, problem-solving, and logical skills essential for everyday life (Bråting & Kilhamn, 2021; Chimoni et al., 2023). Algebra is not merely about mathematical operations but also trains students to think abstractly, recognize patterns, and make generalizations (Newton et al., 2020; Trigueros & Wawro, 2020). Early understanding of algebraic concepts prepares students for more advanced mathematical topics in subsequent educational levels, such as calculus and geometry, which heavily rely on algebraic understanding (Frank & Thompson, 2021; Wasserman et al., 2023). Insufficient algebraic competence can hinder the development of critical thinking, making it difficult to master complex concepts and limiting opportunities in STEM fields and careers (Dolapcioglu & Doğanay, 2022; Spiller et al., 2023). A weak grasp of algebra may also confine students to professions requiring fewer analytical skills, restricting their potential for broader career and life development.

Many students struggle to grasp algebraic concepts, particularly in topics like line equations. Key elements such as slope, intercept, and the general form of linear equations often prove challenging (Flores et al., 2020; Roback & Legler, 2021). Inadequate comprehension of these basic concepts can impact overall algebra understanding (Birgin & Uzun Yazıcı, 2021; Kop et al., 2020). Difficulty connecting graphical representations of lines with their algebraic equations further hinders understanding (Azrillia et al., 2024; Glen & Zazkis, 2021). These challenges are exacerbated by teaching methods that are overly theoretical and lack interactivity or contextual visual exploration. Addressing this issue calls for more interactive and intuitive tools, such as a graphing lines worksheet designed to help students visualize how changes in variables affect line graphs. Visual simulations like PhET Interactive Simulation enable students to experiment and observe the impact of slope and intercept changes on line graphs (Atabas et al., 2020; Oktaviyanthi & Sholahudin, 2023), bridging abstract algebraic concepts with concrete visual representations. To ensure effectiveness, such tools must be evaluated for their impact on algebra competence through research and student feedback. Analyzing the graphing lines worksheet's effectiveness is essential to determine its role in improving student understanding (Gurmu et al., 2024; Pinto & Cañadas, 2021).

To enhance algebra competence, particularly in understanding line equations, integrating Item Response Theory (IRT) into the development of PhET-based graphing lines worksheets is proposed. IRT provides an accurate, individualized measurement of student ability (Kong & Wang, 2021; Wang & Williamson, 2022), enabling analysis of worksheet items by difficulty and discrimination to align materials with student capabilities (Swiecki et al., 2022; Zakwandi et al., 2024). By identifying items that are too challenging or too simple for specific ability levels, IRT supports adaptive learning design (Chan et al., 2021; Murphy et al., 2023). While IRT is well-established in educational measurement, its application to interactive simulation-based worksheets introduces a novel approach tailored to student needs (Cai et al., 2023; Tang & Bao, 2024). This optimization ensures that items challenge students appropriately, fostering meaningful learning across diverse ability levels. Furthermore, IRT-based worksheets provide teachers with detailed student ability profiles, aiding in personalized instruction. IRT also allows for a robust evaluation of worksheet effectiveness, analyzing not just right or wrong answers but also item difficulty and discriminatory power (Alam, 2023; Shultz et al., 2021). By combining interactive PhET simulations with the analytical precision of IRT, these worksheets deliver an engaging learning tool alongside a framework for enhancing algebra competence (Holtom et al., 2022; Wilson, 2023).

Research has widely examined PhET Interactive Simulations across disciplines such as mathematics, physics, and chemistry, highlighting their effectiveness in enhancing conceptual understanding through engaging and intuitive methods (Alhadlaq, 2023; Canoz et al., 2022). In algebra, PhET has proven effective for teaching concepts like line equations, gradients, and constants. PhET has been found to significantly enhance learning outcomes compared to traditional textbook-based methods (Olugbade et al., 2024; Rayan et al., 2023). Similarly, Item Response Theory (IRT) has been extensively applied in educational evaluations, recognized for its precision in measuring student abilities (Kong & Wang, 2021; Wilson, 2023). IRT has been shown to effectively analyze item difficulty and distinguish student abilities (Gan et al., 2020; Stachl & Baranger, 2020). However, combining IRT with interactive tools like PhET for learning evaluations remains rare, presenting a novel avenue for research. This study integrates PhET-based worksheets with IRT-based evaluation, addressing a gap in mathematics education research. By merging these approaches, it aims to enhance algebra learning and provide a robust framework for assessing student competence.

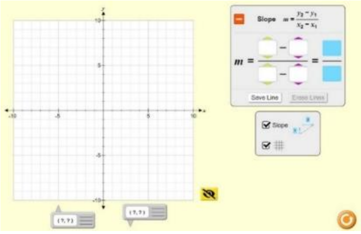
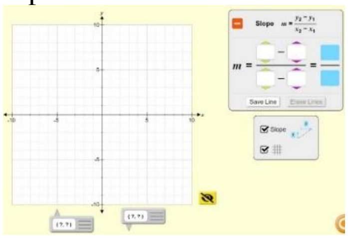
Methods

This study employs a quantitative approach using Item Response Theory (IRT), specifically the two-parameter logistic model (2PL) (König et al., 2019; Riani & Robitzsch, 2024; Wilson, 2023), to analyze the relationship between student abilities and item characteristics in the PhET Interactive Simulation-based Graphing Lines Worksheet, focusing on item difficulty and discrimination. The 2PL model is selected for its ability to evaluate these factors, which are crucial for assessing the effectiveness of the worksheet in supporting students’ understanding of algebraic concepts.

The data analysis was conducted using the Python programming language, executed on the online platfom https://python-fiddle.com/examples/matplotlib. The 2PL IRT model was implemented through Python libraries to calculate individual student ability (theta) and item parameters (difficulty and discrimination). This method enables precise statistical evaluation of the collected data.

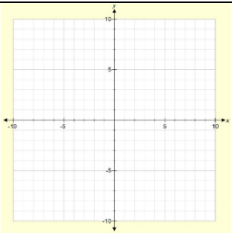
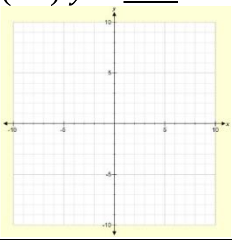
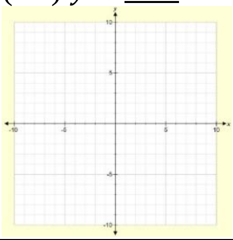
The study involves 120 students completing 12 items on the worksheet, covering four algebra competency indicators: (1) understanding the geometric significance of line slopes, (2) formulating line equations from data or graphs, (3) drawing graphs from line equations, and (4) predicting the effects of variable changes on line equations. The worksheet has a content validity score of ≥0.8 and reliability score of 0.83, confirming its suitability and consistency in measuring students' algebra competencies.

Table 1. The graphing lines worksheet sample items

<table>
<tr><th>Indicator</th><th>Worksheet Sample Items</th></tr>
<tr><td>Understanding the geometric significance of line slopes</td><td>Using the PhET Interactive Simulation, explore the slope of a line directly.
Select values for x_1, x_2, y_1, and y_2 to produce different values of m (slope or gradient).
(a.1) For m = -2, write down the values of x_1, x_2, y_1, and y_2 that you selected, then sketch a visual representation of the line!
</td></tr>
<tr><td>Formulating line equations from data or graphs</td><td>Using the PhET Interactive Simulation, conduct an experiment to determine the equation of a line from two points on a graph.
Calculate the slope and formulate the equation of the line from the given two points.
(a.1) For points (1, 2) and (3, 6), fill in the blanks below!
$m = \frac{y_2 - y_1}{x_2 - x_1}$
$\frac{y - y_1}{x - x_1} = \frac{y_2 - y_1}{x_2 - x_1}$
$\frac{y - \_\_\_\_}{x - \_\_\_\_} = \frac{\_\_\_\_ - \_\_\_\_}{\_\_\_\_ - \_\_\_\_}$
$y - \_\_\_\_ = (x - \_\_\_\_)\left(\frac{\_\_\_\_ - \_\_\_\_}{\_\_\_\_ - \_\_\_\_}\right)$
$y - \_\_\_\_ = (x - \_\_\_\_)(\_\_\_\_)$
$y - \_\_\_\_ = \_\_\_\_x - \_\_\_$
$y = \_\_\_\_x - \_\_\_$
Sketch the visual representation!
</td></tr>
<tr><td>Drawing graphs from line equations</td><td>Using your intuition, knowledge, and learning experience, conduct an experiment to graph the two given line equations.

Draw the graphs of the following two line equations.
(a.1) $y = -x - 4$ dan $y = -x$
At what point does each line intersect the y-axis?
Point (_____, _____)

Does either line intersect the x-axis?</td></tr>
</table>

Indicator	Worksheet Sample Items
	 From the graphed lines, determine: What is the slope of the lines? m = _____ _____________________________ Find the intersection point between the two lines! _____________________________
Predicting the effect of variable changes on line equations	Conduct an experiment by varying the coefficient of *x* and the constant **b** in the given line equation, and observe the interactions that occur. Choose the coefficient values $\boldsymbol{x}$ **= 3**, $\boldsymbol{x}$ **= 1**, $\boldsymbol{x} = \frac{\mathbf{1}}{\mathbf{3}}$, and $\boldsymbol{x} = \frac{\mathbf{1}}{\mathbf{6}}$ in the line equation $y =$ ____$x - 3$, and plot the graphs! (a.1) $y =$ ____$x - 3$  (a.2) $y =$ ____$x - 3$ 

Data is collected through student responses to the 12 worksheet items. Each response is analyzed using the 2PL IRT model to measure individual ability (theta) and item characteristics (difficulty and discrimination) (Ackerman & Ma, 2024; Monroe, 2022; Sweeney et al., 2022). This analysis produces a Person-Item Map, which plots the distribution of student abilities against item difficulties (Kabic & Alexandrowicz, 2023; Scherman & Liebenberg, 2021). This map helps evaluate the effectiveness of the worksheet, especially if student abilities are well-distributed and not concentrated at lower ability levels. Additionally, statistical fit analysis is used to ensure the IRT model's adequacy in representing the collected data (Luong & Flake, 2022; Marsh et al., 2020). The technical data summary matrix for the research can be seen in Table 2 below.

Table 2. Data summary matrix of the research

Analysis Aspects	Description	Key Indicators	Analysis Techniques	Conclusions Drawn
Student Ability Distribution	Analysis of student ability distribution after using the worksheet	- Increase in theta values - Distribution of students across various ability levels	IRT (Estimation of student theta)	If the majority of students are at a high ability level, this indicates that the worksheet is effective in enhancing students' algebraic competency.

Analysis Aspects	Description	Key Indicators	Analysis Techniques	Conclusions Drawn
Item Parameters (Difficulty & Discrimination)	Analysis of item difficulty and discrimination levels in the worksheet	- Item difficulty - Item discrimination	IRT (2PL Model)	If items of high difficulty are answered correctly by students with higher ability, this suggests that the worksheet successfully measures students' algebraic abilities.
Person-Item Map	Mapping of student abilities against item difficulty levels in the worksheet	- Alignment between student abilities and item difficulties	Person-Item Mapping using IRT	If the map shows a well-distributed match between student abilities and item difficulties, the worksheet is deemed appropriate for the students' ability levels.
Fit Statistics	Fit of the IRT model with the collected data	- Model fit statistics values	Fit statistics analysis using IRT	If the IRT model exhibits good fit, it can be concluded that the model accurately represents the data, and the worksheet is considered effective.

Based on Table 2, the analysis reveals several key findings. Student abilities were measured using "theta" values in the IRT 2PL model, where higher theta indicates improved ability. This model, incorporating item discrimination and difficulty, uses binary response data (correct/incorrect) to objectively assess the impact of the worksheet on student abilities. Item parameters, including difficulty and discrimination, were analyzed to evaluate how effectively items distinguish between students with varying abilities. The Person-Item Map was utilized to compare student abilities with item difficulty levels, ensuring the worksheet’s suitability for diverse ability groups. Additionally, statistical fit analysis confirmed that the model accurately represents the data.

Results

Student ability distribution

This study examines the distribution of student ability to evaluate the impact of the PhET Interactive Simulation-based worksheet on algebraic competency. Using the Item Response Theory (IRT) model, it analyzes students' ability levels (theta) and the worksheet's effectiveness across diverse abilities. The histogram below illustrates the student score distribution.

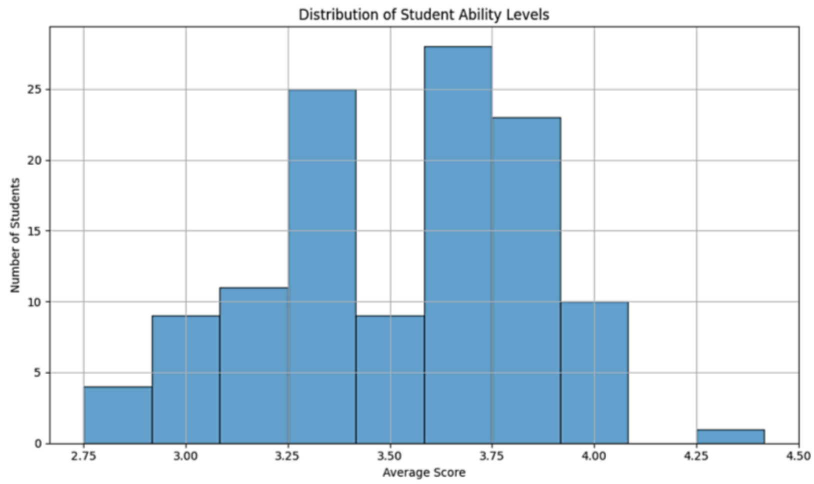


Figure 1. Distribution of student ability levels

The histogram in Figure 1 shows that most students score between 3 and 4, indicating medium ability. Few students score very low (2) or very high (5), suggesting the worksheet effectively differentiates abilities. The distribution is centered, with a slight skew toward higher scores, highlighting the need for targeted interventions to support students with lower scores. Key details of the score distribution are summarized in Table 3.

Tabel 3. Distribution of student score

Point	Aspect	Description
Descriptive Statistics	Mean	The average score of students across all 12 items is 3.43.
	Median	The median score across all students is 4.
	Variance	The dispersion of scores from the mean is 0.78.
	Standard Deviation	The standard deviation of student scores is 0.88.
Score Distribution	Score Range	The range of student scores extends from 2 (lowest) to 5 (highest).
	Score Frequency	Score 2: 16% of total scores Score 3: 32% of total scores Score 4: 34% of total scores Score 5: 18% of total scores

The analysis of student ability distribution based on responses to the 12 worksheet items (Table 3) reveals an average score of 3.43 and a median of 4. With a variance of 0.78 and a standard deviation of 0.88, the scores show moderate variation, indicating that most students have relatively uniform abilities in solving the problems. Figure 2 visually represents student response patterns to the graphing lines worksheet items.

Figure 2 illustrates score patterns for 120 students (S1-S120) across the 12 items in the graphing lines worksheet. Light colors (e.g., red) indicate higher scores or “easier items,” while dark colors (e.g., blue) signify lower scores or “more challenging items.” Numerical values within the colored boxes enhance clarity. The heatmap reveals patterns: item 2 shows predominantly dark colors, indicating high difficulty, while item 12 is mostly red, signifying ease. Items 5 and 6 also appear easier, whereas item 11, with darker tones, reflects higher difficulty.

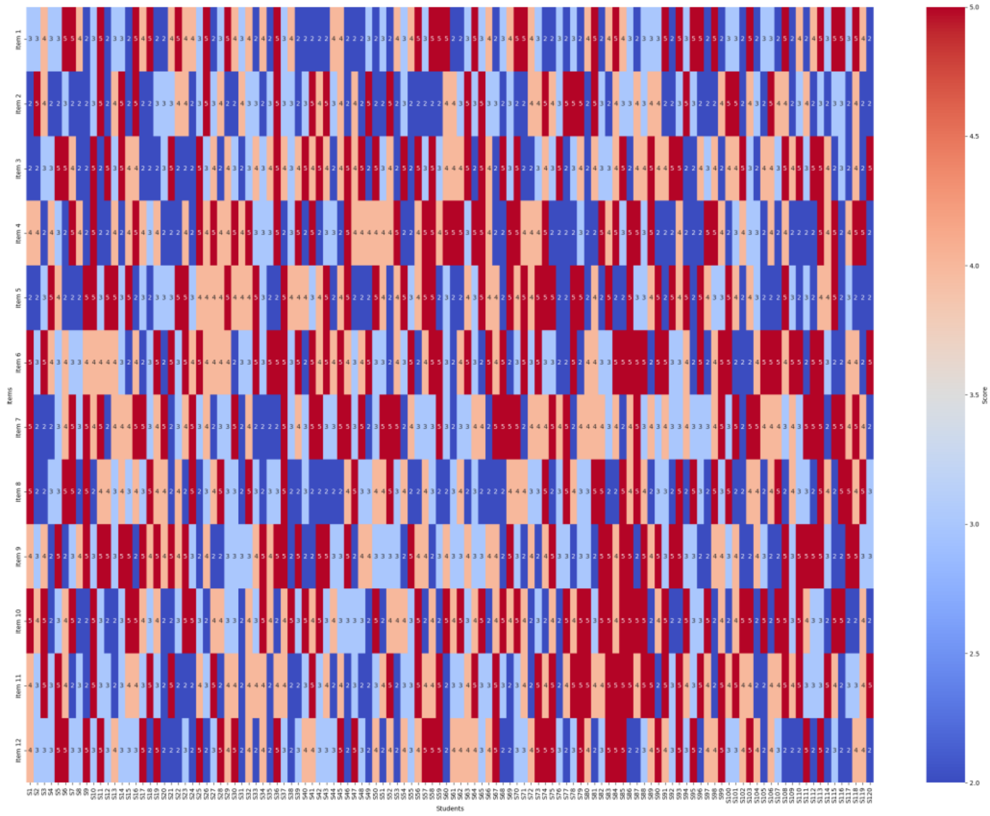


Figure 2. Heatmap of student scores per item

Item parameters

Using the 2-Parameter Logistic (2PL) model in Item Response Theory (IRT), item parameters provide insights into the worksheet's characteristics. The 2PL model assesses item difficulty and discrimination, analyzing how well each item differentiates students based on their abilities. The Item Characteristic Curve (ICC) visually represents item performance and differentiation across ability levels. Figure 3 displays the ICC for the 12 worksheet items based on student responses.

The Item Characteristic Curve (ICC) in Figure 3 shows the relationship between students' ability levels (theta) and the probability of correctly answering an item. The curve highlights item response patterns across varying ability levels. A steep ICC indicates high discrimination, effectively distinguishing between lower- and higher-ability students, while a flatter curve signifies lower discrimination. The x-axis represents student ability (theta), and the y-axis shows the probability of a correct response. The curve's position along the x-axis reflects item difficulty: items further right is more challenging, while those to the left are easier. Appropriately challenging items, correctly answered by higher-ability students, demonstrate alignment with skill measurement goals.

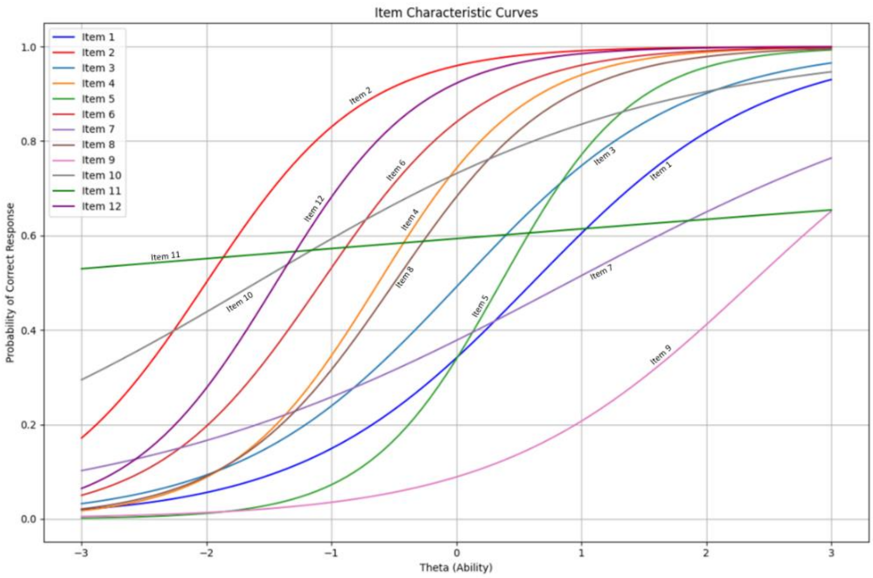


Figure 3. Item characteristic curve

In the IRT model, item difficulty is defined as the ability level at which students have a 50% probability of answering correctly. Higher difficulty values indicate more challenging items. For the PhET-based graphing lines worksheet, high-difficulty items require a deeper grasp of algebraic concepts. Figure 4 displays the distribution of student responses by item difficulty for this worksheet.

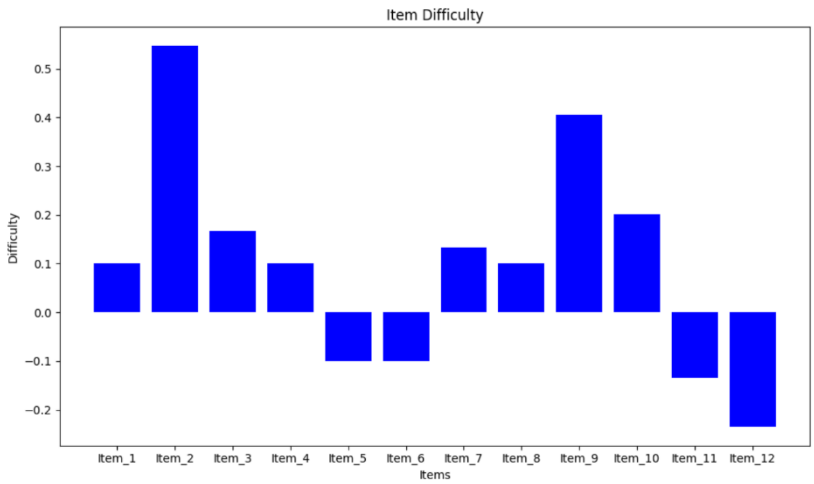


Figure 4. Item difficulty in graphing lines worksheet

Figure 4 illustrates the distribution of difficulty levels for each item in the graphing lines worksheet. The x-axis represents the items, while the y-axis shows difficulty levels, ranging from negative (easier) to positive (more difficult). Positive values indicate items correctly answered primarily by higher-ability students, whereas negative values correspond to items that are easier for lower-ability students. This distribution helps assess whether the test items appropriately cover a range of difficulty levels for evaluating student abilities.

Item discrimination, which reflects how effectively an item distinguishes between students with high and low abilities, is also analyzed. Items with high discrimination better identify students with strong conceptual understanding, while low-discrimination items provide limited differentiation. Figure 5 visualizes the item discrimination values from this analysis.

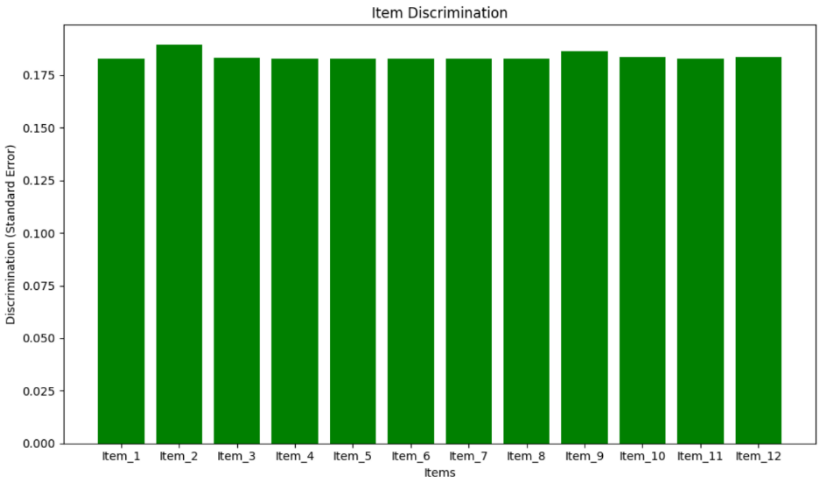


Figure 5. Item discrimination in graphing lines worksheet

Figure 5 presents the discrimination values for each item, indicating how well an item differentiates between students with high and low abilities. The x-axis represents item identification, and the y-axis shows the discrimination values. The even distribution and sufficient discrimination values across items suggest that the graphing lines worksheet is fair and effective in evaluating student abilities. All items have similar discrimination values, around 0.175, meaning each item is effective at distinguishing between students with varying ability levels. No items exhibit low discrimination, ensuring that all items accurately measure differences in student abilities.

The results of the item analysis, including the ICC curves in Figure 3, the difficulty distribution in Figure 4, and the discrimination values in Figure 5, are summarized in Table 4.

Table 4. Item analysis of graphing lines worksheet

Item	Difficulty	Discrimination	Description
1	0.608262	1.083542	Item 1 has a moderate difficulty of 0.608 and good discrimination (1.08), making it effective for distinguishing intermediate-level students.
2	-2.003021	1.581835	Item 2, with a difficulty of -2.00 and high discrimination (1.58), is easy and well-suited for differentiating lower-ability students.
3	0.031188	1.121797	Item 3, with a difficulty of 0.03 and discrimination of 1.12, is moderately difficult and effective for intermediate-level students.
4	-0.623778	1.696907	Item 4, with a difficulty of -0.62 and high discrimination (1.69), is easy and distinguishes students from low to intermediate abilities.
5	0.355491	1.882018	Item 5, with a difficulty of 0.36 and discrimination of 1.88, is moderately difficult with high discrimination, suitable for intermediate to high-ability students.
6	-1.082004	1.535193	Item 6, with a difficulty of -1.08 and high discrimination (1.54), is easy and effective for differentiating lower-ability students.
7	0.892322	0.558119	Item 7, with a difficulty of 0.89 and low discrimination (0.56), is challenging but not effective

Item	Difficulty	Discrimination	Description
			in distinguishing ability levels, suggesting it may need review.
8	-0.497054	1.535308	Item 8, with a difficulty of -0.50 and discrimination of 1.53, is easy and effective for differentiating intermediate-ability students.
9	2.363910	0.984728	Item 9, with a difficulty of 2.36 and moderate discrimination (0.98), is the most difficult and suitable for high-ability students.
10	-1.602443	0.624820	Item 10, with a difficulty of -1.60 and low discrimination (0.62), is easy but ineffective at distinguishing abilities.
11	-4.378166	0.086401	Item 11, with a difficulty of -4.38 and very low discrimination (0.086), is the easiest but not useful for differentiating abilities.
12	-1.442709	1.718028	Item 12, with a difficulty of -1.44 and high discrimination (1.72), is easy and well-suited for assessing abilities from low to intermediate levels.

Person-item map

The Person-Item Map illustrates the distribution of participant abilities and item difficulty levels on the logit scale. Higher-ability participants are placed at the top, while lower-ability participants are positioned at the bottom. More difficult items are placed to the right, with easier items on the left. Figure 6 below shows the person-item map for 120 participants and the 12 items of the graphing lines worksheet.

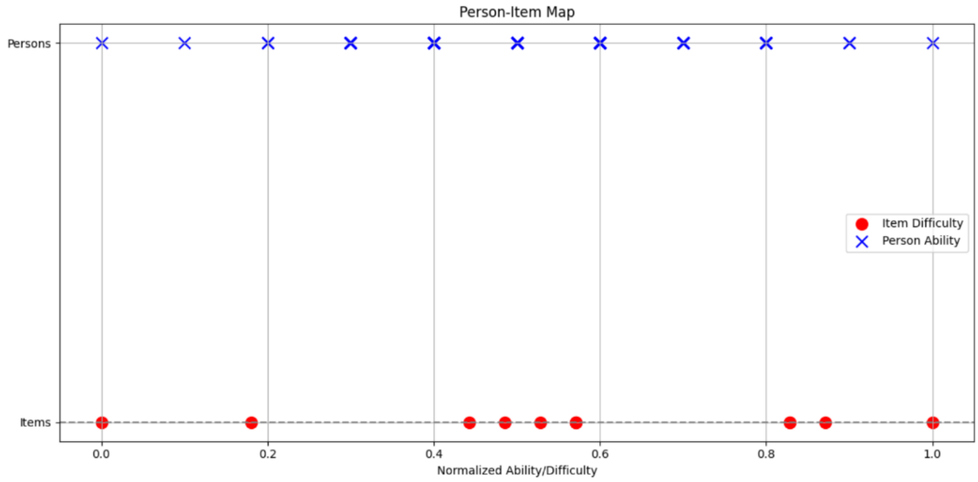


Figure 6. Distribution of participant abilities and item difficulty levels on the logit scale

From Figure 6 and the person-item map analysis, it is clear that participant abilities and item difficulties are represented on the X and Y axes using the logit scale. The X-axis shows both participant abilities and item difficulties, where higher logit values for participants indicate greater abilities, and higher logit values for items signify greater difficulty. The Y-axis is split into two levels: Level 0 for item positions and Level 1 for participant positions, with a horizontal reference line at value 0. Red dots represent item positions on the difficulty scale, while blue dots indicate participant abilities. The further a red dot is from the centerline, the more difficult the item; the further a blue dot is from the centerline, the higher or lower the participant's ability.

The data is normalized to enable comparison between participant abilities and item difficulties. Participants and items are classified into three categories: high-ability participants and difficult items (logit ≥ 1), consisting of 26 participants and no items; medium-ability participants and medium-difficulty items (logit 0 to 1), comprising 37 participants and 8 items; and low-ability participants and easy items (logit < 0), including 57 participants and 4 items. Details of the worksheet items based on logit values are provided in Table 5.

Table 5 shows the item categories and subject completion based on logit values in the worksheet. The logit values categorize the items into two groups: "Moderate" for items with positive logit values and "Easy" for items with negative logit values. The table includes 12 items, with items 1 to 10 categorized as moderate, and items 5, 6, 11, and 12 classified as easy.

Tabel 5. Item categories and subject completion

Item Number	Logit Value	Item Category	Item Number	Logit Value	Item Category
1	0.1001	Moderate	7	0.1335	Moderate
2	0.5465	Moderate	8	0.1001	Moderate
3	0.1671	Moderate	9	0.4055	Moderate
4	0.1001	Moderate	10	0.2007	Moderate
5	-0.1001	Easy	11	-0.1335	Easy
6	-0.1001	Easy	12	-0.2344	Easy

Fit statistics

Fit statistics are used to assess how well the IRT model matches the data. If these statistics indicate a good fit, it suggests that the worksheet accurately reflects student abilities and item difficulties. Figure 7 presents the results of the fit statistics analysis for the research data.

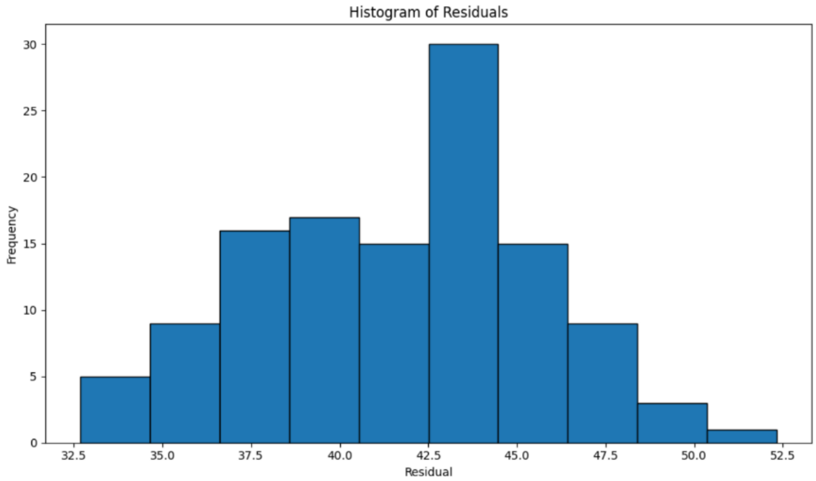


Figure 7. Distribution of fit statistics

In Figure 7, the residual distribution can be interpreted through the shape and skewness of the data. The histogram shape indicates that the residual distribution approximates a normal distribution, with the center around a residual value of 42.5. Most residuals are concentrated between 40 and 45, peaking at approximately 30 residuals at 42.5, indicating that the model's predictions are generally close to the observed values. Some residuals are further spread out on both sides, but no significant outliers are present, suggesting that while the model is mostly accurate, there are a few instances where predictions do not match actual data. The residual distribution shows a slight positive skew, with a tail to the right, confirming that the IRT model generally fits the data well.

Discussion

This study evaluated students' ability distribution and the effectiveness of a PhET Interactive Simulation-based graphing lines worksheet using the Item Response Theory (IRT) approach. The findings showed that most students demonstrated moderate ability, with an average score of 3.43 and a median of 4, indicating competence in solving items of moderate to high difficulty. Analysis of responses to 12 worksheet items revealed that while most were answered correctly, items 2 and 11 posed significant challenges, requiring deeper understanding. Heatmap analysis identified item 12 as the easiest and item 2 as the most difficult. Using the 2-Parameter Logistic Model (2PL), items displayed good discrimination, with steep item characteristic curves effectively distinguishing between low- and high-ability students. These results affirm the worksheet’s effectiveness in assessing diverse student abilities and reinforcing key algebraic concepts like graphing lines.

This study reinforces previous findings on the effectiveness of PhET Interactive Simulations in enhancing student competency while highlighting the role of simulation-based worksheets as strategic learning tools, particularly for graphing lines. The worksheet was designed to build conceptual understanding of slope, linear equations, and the impact of variable changes on graphs. Interactive simulations have been shown to improve abstract conceptual understanding and deepen engagement (Huang et al., 2022; Sugden et al., 2021). This study extends these findings by incorporating a structured worksheet that guides students to utilize the simulation effectively. The integration of the worksheet with the PhET Interactive Simulation provided a systematic approach to applying algebraic concepts. Students engaged with the simulation independently while being guided through activities designed to: (1) explore slope geometrically, (2) form linear equations from data or graphs, and (3) predict graph changes resulting from coefficient or constant variations.

A key contribution of this study lies in combining the graphing lines worksheet with Item Response Theory (IRT) and the 2-Parameter Logistic Model (2PL) to analyze students’ responses (Scherman & Liebenberg, 2021). This analysis evaluates item difficulty and discrimination, positioning the worksheet as both a learning tool and a robust evaluation instrument. Quantitative insights into item parameters highlight students’ conceptual understanding and how difficulty levels influence performance (Gan et al., 2020; Sweeney et al., 2022). Moreover, this study demonstrates that combining PhET Interactive Simulations with structured worksheets effectively addresses varying student ability levels. Item parameter analysis aids in designing questions that challenge students appropriately, catering to both lower and higher abilities (König et al., 2019). The worksheet thus serves as both a learning resource and a diagnostic tool, identifying difficulties and enabling question adjustments to meet diverse needs.

This study uniquely evaluates simulation-based learning in Indonesia, particularly the development of PhET Interactive Simulation-based mathematics worksheets integrated with the IRT model. Its novelty lies in quantitatively analyzing student ability distribution using logit scales and item parameters through the 2PL model, as well as introducing the rarely used Person-Item Map in secondary education. Findings reveal that logit-based analysis and statistical fit provide precise insights into the alignment of student abilities with item difficulty, paving the way for further application of the IRT approach in technology-enhanced learning.

Key recommendations include: (1) revising challenging items, such as items 2 and 11, to better align with student abilities while maintaining appropriate difficulty; (2) integrating additional interactive simulations to support more advanced concepts; (3) implementing targeted interventions for lower-ability students identified through logit-based analysis; and (4) conducting trials in schools with diverse demographics to expand the worksheet's applicability.

Limitations of the study include: (1) the sample size of 120 students may not fully capture the diversity of student abilities; (2) IRT-based analysis requires larger samples for stable item parameters, necessitating future studies with broader populations; (3) the focus on graphing lines limits generalizability to other algebraic concepts; and (4) external factors like technology availability and school infrastructure were not examined. Addressing these limitations in future research could yield deeper insights into the effectiveness of simulation-based worksheets for enhancing algebraic abilities.

Conclusion

The integration of Item Response Theory (IRT) in the development of a graphing lines worksheet using PhET Interactive Simulation effectively enhances students' algebraic competence. The application of IRT, particularly the 2-Parameter Logistic (2PL) model, enabled a deeper analysis of students' abilities in responding to the worksheet items, while also identifying the difficulty level and discrimination power of each item. Item 2 was found to be the most difficult, warranting further review to ensure its alignment with the intended measurement objectives. In contrast, Item 12 was deemed too easy and may not provide significant insights into higher-ability students. Meanwhile, Items 1 and 11 exhibited balanced difficulty levels and effective discrimination, making them robust items for evaluating student performance. The heatmap revealed variations in student response patterns, indicating that the test is capable of assessing students with a wide range of abilities. The Item Characteristic Curve (ICC) analysis further reinforced the findings related to item difficulty and discrimination, offering deeper insights into how each item functions within the measurement framework.

Overall, the combination of interactive simulation and the worksheet provided a more structured approach for students to learn graphical algebra concepts, such as line slopes and linear equations, thereby enhancing their conceptual understanding. The study recommends that the development of PhET-based worksheets be expanded to include other algebraic topics, such as quadratic equations and systems of linear equations, to further improve the effectiveness of learning. Additionally, further review of items that are too easy or difficult is necessary to ensure they align with students' abilities. Moreover, the application of IRT methods to other subjects and more diverse populations could significantly contribute to creating more adaptive evaluation instruments.

Acknowledgment

We gratefully acknowledge the Ministry of Education, Culture, Research, and Technology (Kemendikbud Ristek) for funding this research. Special thanks to our academic advisors for their guidance and to the 120 middle school students in Serang who participated as respondents. We also appreciate those who assisted with proofreading, data processing, and logistics, making this study possible.

Conflicts of Interest

The authors affirm that there are no conflicts of interest related to the publication of this manuscript. Furthermore, all ethical considerations have been adhered to, including those regarding plagiarism, research misconduct, data fabrication or falsification, duplicate publication or submission, and unnecessary redundancies.

Funding Statement

This work was supported by the Ministry of Education, Culture, Research, and Technology of Indonesia [contract number 106/E5/PG.02.00.PL/2024, dated June 11, 2024]; LLDIKTI Region IV [contract number 038/SP2H/RT-MONO/LL4/2024, dated June 14, 2024]; and Universitas Serang Raya [contract number 003/DRTPM.PDPREGULER/UNSERA/VI/2024, dated June 19, 2024]

Author Contributions

Giyanti: Conceptualization, resources, writing - original draft, and visualization; **Indri Lestari:** Project administration, validation, software and supervision; **Rina Oktaviyanthi:** Data curation, formal analysis, and methodology, writing - review and editing

References

Ackerman, T. A., & Ma, Y. (2024). Examining differential item functioning from a multidimensional IRT perspective. *Psychometrika*, *89*(1), 4–41. https://doi.org/10.1007/s11336-024-09965-6

Alam, A. (2023). Harnessing the power of AI to create intelligent tutoring systems for enhanced classroom experience and improved learning outcomes. *Lecture Notes on Data Engineering and Communications Technologies*, *171*, 571–591. https://doi.org/10.1007/978-981-99-1767-9_42

Alhadlaq, A. (2023). Computer-based simulated learning activities: exploring saudi students' attitude and experience of using simulations to facilitate unsupervised learning of science concepts. *Applied Sciences 2023, Vol. 13, Page 4583*, *13*(7), 4583. https://doi.org/10.3390/app13074583

Atabas, S., Schellinger, J., Whitacre, I., Findley, K., & Hensberry, K. (2020). A tale of two sets of norms: Comparing opportunities for student agency in mathematics lessons with and without interactive simulations. *The Journal of Mathematical Behavior*, *58*, 100761. https://doi.org/10.1016/j.jmathb.2020.100761

Azrillia, W., Oktaviyanthi, R., Khotimah, K., & Garcia, M. L. B. (2024). Feasibility test of articulate storyline 3 learning media based on local wisdom for optimizing students' algebraic thinking skills. *Paedagogia*, *27*(1), 1–15. https://doi.org/10.20961/paedagogia.v27I1.84145

Birgin, O., & Uzun Yazıcı, K. (2021). The effect of GeoGebra software–supported mathematics instruction on eighth-grade students' conceptual understanding and retention. *Journal of Computer Assisted Learning*, *37*(4), 925–939. https://doi.org/10.1111/jcal.12532

Bråting, K., & Kilhamn, C. (2021). Exploring the intersection of algebraic and computational thinking. *Mathematical Thinking and Learning*, *23*(2), 170–185. https://doi.org/10.1080/10986065.2020.1779012

Cai, L., Chung, S. W., & Lee, T. (2023). Incremental model fit assessment in the case of categorical data: Tucker–Lewis index for item response theory modeling. *Prevention Science*, *24*(3), 455–466. https://doi.org/10.1007/s11121-021-01253-4

Canoz, G. M., Ucar, S., & Demircioglu, T. (2022). Investigate the effect of argumentation-promoted interactive simulation applications on students' argumentation levels, academic achievements, and entrepreneurship skills in science classes. *Thinking Skills and Creativity*, *45*, 101106. https://doi.org/10.1016/j.tsc.2022.101106

Chan, S. W., Looi, C. K., & Sumintono, B. (2021). Assessing computational thinking abilities among Singapore secondary students: a Rasch model measurement analysis. *Journal of Computers in Education*, *8*(2), 213–236. https://doi.org/10.1007/s40692-020-00177-2

Chimoni, M., Pitta-Pantazi, D., & Christou, C. (2023). Unfolding algebraic thinking from a cognitive perspective. *Educational Studies in Mathematics*, *114*(1), 89–108. https://doi.org/10.1007/S10649-023-10218-z

Dolapcioglu, S., & Doğanay, A. (2022). Development of critical thinking in mathematics classes via authentic learning: an action research. *International Journal of Mathematical Education in Science and Technology*, *53*(6), 1363–1386. https://doi.org/10.1080/0020739X.2020.1819573

Flores, C. D., López, M. I. R., & Moore-Russo, D. (2020). Conceptualizations of slope in Mexican intended curriculum. *School Science and Mathematics*, *120*(2), 104–115. https://doi.org/10.1111/ssm.12389

Frank, K., & Thompson, P. W. (2021). School students' preparation for calculus in the United States. *ZDM - Mathematics Education*, *53*(3), 549–562. https://doi.org/10.1007/S11858-021-01231-8

Gan, W., Sun, Y., Peng, X., & Sun, Y. (2020). Modeling learner's dynamic knowledge construction procedure and cognitive item difficulty for knowledge tracing. *Applied Intelligence*, *50*(11), 3894–3912. https://doi.org/10.1007/S10489-020-01756-7

Glen, L., & Zazkis, R. (2021). On linear functions and their graphs: refining the cartesian connection. *International Journal of Science and Mathematics Education*, *19*(7), 1485–1504. https://doi.org/10.1007/S10763-020-10113-6

Gurmu, F., Tuge, C., & Hunde, A. B. (2024). Effects of GeoGebra-assisted instructional methods on students' conceptual understanding of geometry. *Cogent Education*, *11*(1). https://doi.org/10.1080/2331186X.2024.2379745

Holtom, B., Baruch, Y., Aguinis, H., & A Ballinger, G. (2022). Survey response rates: Trends and a validity assessment framework. *Human Relations*, *75*(8), 1560–1584. https://doi.org/10.1177/00187267211070769

Huang, Y. M., Silitonga, L. M., & Wu, T. T. (2022). Applying a business simulation game in a flipped classroom to enhance engagement, learning achievement, and higher-order thinking skills. *Computers & Education*, *183*, 104494. https://doi.org/10.1016/j.compedu.2022.104494

Kabic, M., & Alexandrowicz, R. W. (2023). RMX/PIccc: An Extended Person–Item Map and a Unified IRT Output for eRm, psychotools, ltm, mirt, and TAM. *Psych 2023, Vol. 5, Pages 948-965*, *5*(3), 948–965. https://doi.org/10.3390/psych5030062

Kong, S. C., & Wang, Y. Q. (2021). Item response analysis of computational thinking practices: Test characteristics and students' learning abilities in visual programming contexts. *Computers in Human Behavior*, *122*, 106836. https://doi.org/10.1016/j.chb.2021.106836

König, C., Spoden, C., & Frey, A. (2019). An optimized bayesian hierarchical two-parameter logistic model for small-sample item calibration. *Applied Psychological Measurement*, *44*(4), 311–326. https://doi.org/10.1177/0146621619893786

Kop, P. M. G. M., Janssen, F. J. J. M., Drijvers, P. H. M., & van Driel, J. H. (2020). The relation between graphing formulas by hand and students' symbol sense. *Educational Studies in Mathematics*, *105*(2), 137–161. https://doi.org/10.1007/s10649-020-09970-3

Luong, R., & Flake, J. K. (2022). Measurement invariance testing using confirmatory factor analysis and alignment optimization: A Tutorial for transparent analysis planning and reporting. *Psychological Methods*, *28*(4), 905–924. https://doi.org/10.1037/met0000441

Marsh, H. W., Guo, J., Dicke, T., Parker, P. D., & Craven, R. G. (2020). confirmatory factor analysis (CFA), exploratory structural equation modeling (ESEM), and set-ESEM: Optimal balance between goodness of fit and parsimony. *Multivariate Behavioral Research*, *55*(1), 102–119. https://doi.org/10.1080/00273171.2019.1602503

Monroe, S. (2022). Item response theory. *Item Response Theory*. https://doi.org/10.4324/9781138609877-ree61-1

Murphy, D. H., Little, J. L., & Bjork, E. L. (2023). The value of using tests in education as tools for learning—not just for assessment. *Educational Psychology Review*, *35*(3), 1–21. https://doi.org/10.1007/s10648-023-09808-3

Newton, K. J., Barbieri, C. A., & Booth, J. L. (2020). Key mathematical competencies from arithmetic to algebra. *Oxford Research Encyclopedia of Education*. https://doi.org/10.1093/acrefore/9780190264093.013.956

Oktaviyanthi, R., & Sholahudin, U. (2023). Phet assisted trigonometric worksheet for students' trigonometric adaptive thinking. *Mosharafa: Jurnal Pendidikan Matematika*, *12*(2), 229–242. https://doi.org/10.31980/mosharafa.v12I2.779

Olugbade, D., Oyelere, S. S., & Agbo, F. J. (2024). Enhancing junior secondary students' learning outcomes in basic science and technology through PhET: A study in Nigeria. *Education and Information Technologies*, *29*(11), 14035–14057. https://doi.org/10.1007/S10639-023-12391-3

Pinto, E., & Cañadas, M. C. (2021). Generalizations of third and fifth graders within a functional approach to early algebra. *Mathematics Education Research Journal*, *33*(1), 113–134. https://doi.org/10.1007/S13394-019-00300-2

Rayan, B., Daher, W., Diab, H., & Issa, N. (2023). Integrating PhET simulations into elementary science education: A qualitative analysis. *Education Sciences 2023, Vol. 13, Page 884*, *13*(9), 884. https://doi.org/10.3390/educsci13090884

Riani, M., & Robitzsch, A. (2024). Estimation of standard error, linking error, and total error for robust and nonrobust linking methods in the two-parameter logistic model. *Stats*, *7, 592-612*, *7*(3), 592–612. https://doi.org/10.3390/stats7030036

Roback, P., & Legler, J. (2021). Beyond multiple linear regression: Applied generalized linear models and multilevel models in R. *Beyond Multiple Linear Regression*. https://doi.org/10.1201/9780429066665

Scherman, V., & Liebenberg, L. (2021). Item response theory integrating qualitative data. *The Routledge Reviewer's Guide to Mixed Methods Analysis*, 117–123. https://doi.org/10.4324/9780203729434

Shultz, K. S. ., Whitney, D. J. ., & Zickar, M. J. . (2021). *Measurement theory in action : case studies and exercises*. Routledge, Taylor and Francis Group. https://www.routledge.com/Measurement-Theory-in-Action-Case-Studies-and-Exercises/Shultz-Whitney-Zickar/p/book/9780367192181

Spiller, J., Clayton, S., Cragg, L., Johnson, S., Simms, V., & Gilmore, C. (2023). Higher level domain specific skills in mathematics; The relationship between algebra, geometry, executive function skills and mathematics achievement. *PLOS ONE*, *18*(11), e0291796. https://doi.org/10.1371/journal.pone.0291796

Stachl, C. N., & Baranger, A. M. (2020). Sense of belonging within the graduate community of a research-focused STEM department: Quantitative assessment using a visual narrative and item response theory. *PLOS ONE*, *15*(5), e0233431. https://doi.org/10.1371/journal.pone.0233431

Sugden, N., Brunton, R., MacDonald, J. B., Yeo, M., & Hicks, B. (2021). Evaluating student engagement and deep learning in interactive online psychology learning activities. *Australasian Journal of Educational Technology*, *37*(2), 45–65. https://doi.org/10.14742/ajet.6632

Sweeney, S. M., Sinharay, S., Johnson, M. S., & Steinhauer, E. W. (2022). An Investigation of the Nature and Consequence of the Relationship between IRT Difficulty and Discrimination. *Educational Measurement: Issues and Practice*, *41*(4), 50–67. https://doi.org/10.1111/emip.12522

Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., Selwyn, N., & Gašević, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence*, *3*, 100075. https://doi.org/10.1016/j.caeai.2022.100075

Tang, H., & Bao, Y. (2024). Self-regulated learner profiles in MOOCs: A cluster analysis based on the item response theory. *Interactive Learning Environments*. https://doi.org/10.1080/10494820.2022.2129394

Trigueros, M., & Wawro, M. (2020). Linear algebra teaching and learning. *Encyclopedia of Mathematics Education*, 474–478. https://doi.org/10.1007/978-3-030-15789-0_100021

Wang, G., & Williamson, A. (2022). Course evaluation scores: valid measures for teaching effectiveness or rewards for lenient grading? *Teaching in Higher Education*, *27*(3), 297–318. https://doi.org/10.1080/13562517.2020.1722992

Wasserman, N. H., Buchbinder, O., & Buchholtz, N. (2023). Making university mathematics matter for secondary teacher preparation. *ZDM - Mathematics Education*, *55*(4), 719–736. https://doi.org/10.1007/s11858-023-01484-5

Wilson, Mark. (2023). *Constructing measures: An item response modeling approach*. Routledge. https://www.routledge.com/Constructing-Measures-An-Item-Response-Modeling-Approach/Wilson/p/book/9781032261683

Zakwandi, R., Istiyono, E., & Dwandaru, W. S. B. (2024). A two-tier computerized adaptive test to measure student computational thinking skills. *Education and Information Technologies*, *29*(7), 8579–8608. https://doi.org/10.1007/s10639-023-12093-w